Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis

Samra Turajlic, Kevin Litchfield*, Hang Xu, Rachel Rosenthal, Nicholas McGranahan, James L Reading, Yien Ning S Wong, Andrew Rowan, Nnennaya Kanu, Maise Al Bakir, Tim Chambers, Roberto Salgado, Peter Savas, Sherene Loi, Nicolai J Birkbak, Laurent Sansregret, Martin Gore, James Larkin, Sergio A Quezada, Charles Swanton*

Summary

Background The focus of tumour-specific antigen analyses has been on single nucleotide variants (SNVs), with the contribution of small insertions and deletions (indels) less well characterised. We investigated whether the frameshift nature of indel mutations, which create novel open reading frames and a large quantity of mutagenic peptides highly distinct from self, might contribute to the immunogenic phenotype.

Methods We analysed whole-exome sequencing data from 5777 solid tumours, spanning 19 cancer types from The Cancer Genome Atlas. We compared the proportion and number of indels across the cohort, with a subset of results replicated in two independent datasets. We assessed in-silico tumour-specific neoantigen predictions by mutation type with pan-cancer analysis, together with RNAseq profiling in renal clear cell carcinoma cases (n=392), to compare immune gene expression across patient subgroups. Associations between indel burden and treatment response were assessed across four checkpoint inhibitor datasets.

Findings We observed renal cell carcinomas to have the highest proportion (0·12) and number of indel mutations across the pan-cancer cohort ($p<2·2\times10^{-16}$), more than double the median proportion of indel mutations in all other cancer types examined. Analysis of tumour-specific neoantigens showed that enrichment of indel mutations for high-affinity binders was three times that of non-synonymous SNV mutations. Furthermore, neoantigens derived from indel mutations were nine times enriched for mutant specific binding, as compared with non-synonymous SNV derived neoantigens. Immune gene expression analysis in the renal clear cell carcinoma cohort showed that the presence of mutant-specific neoantigens was associated with upregulation of antigen presentation genes, which correlated (r=0·78) with T-cell activation as measured by CD8-positive expression. Finally, analysis of checkpoint inhibitor response data revealed frameshift indel count to be significantly associated with checkpoint inhibitor response across three separate melanoma cohorts ($p=4·7\times10^{-4}$).

Interpretation Renal cell carcinomas have the highest pan-cancer proportion and number of indel mutations. Evidence suggests indels are a highly immunogenic mutational class, which can trigger an increased abundance of neoantigens and greater mutant-binding specificity.

Funding Cancer Research UK, UK National Institute for Health Research (NIHR) at the Royal Marsden Hospital National Health Service Foundation Trust, Institute of Cancer Research and University College London Hospitals Biomedical Research Centres, the UK Medical Research Council, the Rosetrees Trust, Novo Nordisk Foundation, the Prostate Cancer Foundation, the Breast Cancer Research Foundation, the European Research Council.

Introduction

Tumour mutations are a key substrate for the generation of anticancer immunity.[1] Large-scale sequencing studies[2] have led to the systematic annotation of mutational processes and somatic alterations across a broad range of human cancer types. The cumulative insight from these studies has advanced our understanding of oncogenesis at both a basic and translational level. The data have also been scrutinised for mutations that might play a role in the recognition of cancer cells by the immune system. The focus of these analyses to a large extent is on the single nucleotide variants (SNVs), on account of the relative simplicity and reliability of calling sequence changes of one base pair (bp) fixed length. As a consequence, the effect of small scale insertion and deletion mutations (indels) on antitumour immunity has been poorly characterised despite the clear link of such mutations to oncogenesis[3] and their potential to generate highly immunogenic peptides.

The success of checkpoint inhibitor therapies underlines the notion that tumour-specific T-cell responses pre-exist in some patients and are kept under tight control via immune modulatory mechanisms. To date, checkpoint inhibitors have been approved for the treatment of six solid tumour types: melanoma (anti-PD-1/CTLA-4), merkel cell carcinoma (anti-PDL-1), renal clear cell carcinoma (anti-PD-1), non-small cell lung cancer (lung adenocarcinoma and lung squamous cell

Lancet Oncol 2017

Published Online July 7, 2017 http://dx.doi.org/10.1016/S1470-2045(17)30516-8

*Joint first authors

Translational Cancer Therapeutics Laboratory, The Francis Crick Institute, London, UK (S Turajlic MD, K Litchfield PhD, H Xu PhD, N McGranahan PhD, A Rowan BSc, M Al Bakir MD, T Chambers MSc, N J Birkbak PhD, L Sansregret PhD, Prof C Swanton FRCP)**; Renal and Skin Units, The Royal Marsden Hospital National Health Service Foundation Trust, London, UK** (S Turajlic, Prof M Gore FRCP, J Larkin FRCP)**; Cancer Immunology Unit, Research Department of Haematology** (J L Reading PhD, Y N S Wong MD, S A Quezada) **and Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, Paul O'Gorman Building, London, UK** (R Rosenthal MSc, N Kanu PhD, Prof C Swanton, N McGranahan)**; Department of Pathology, Gasthuiszusters, Antwerp, Belgium** (R Salgado MD)**; Division of Research and Cancer Medicine, Peter MacCallum Cancer Centre, University of Melbourne, Melbourne, VIC, Australia** (P Savas MD, S Loi MD, R Salgado)**; and Department of Medical Oncology, University College London Hospitals, London, UK** (Prof C Swanton)

Correspondence to: Prof Charles Swanton, Translational Cancer Therapeutics Laboratory, The Francis Crick Institute, London NW1 1AT, UK **charles.swanton@crick.ac.uk**

Research in context

Evidence before this study

We searched for available evidence in PubMed, which revealed multiple publications documenting overall mutation rates and signatures by cancer type. The predominant focus of existing literature was on single nucleotide variation (SNV) mutations, with no previous study done of insertion and deletion (indel) mutations on a pan-cancer basis. Regarding the association between somatic mutations and upregulation of antitumour immunity via checkpoint inhibition, several previous studies reported a link between high SNV load and improved response to checkpoint inhibition. Prevailing evidence suggests the mechanism of this association is linked to tumour-specific neoantigen reactive T cells. No previous pan-cancer study has investigated the difference between SNV and indel-derived neoantigens, despite the propensity of indels to generate highly mutagenic peptides via creation of a shifted novel open reading frame.

Added value of this study

We did a pan-cancer assessment of indel load across 5777 tumour samples spanning 19 cancer types. Kidney tumours were observed to have the highest proportion and absolute count of indel mutations on a pan-cancer basis, a result which was replicated in two further independent datasets. Compared with SNV mutations, indel mutations were observed to generate three times more high-binding-affinity neoantigens, and nine times more mutant-specific binders. Finally, we assessed the association between indel load and checkpoint inhibitor response in three melanoma cohorts, which showed indel load to be more strongly associated with response than non-synonymous (ns) SNV load.

Implications of all the available evidence

Our data highlight the importance of frameshift neoantigens alongside nsSNV neoantigens as determinants of immunotherapy efficacy and potentially crucial targets for vaccine and cell therapy interventions. Our observations in kidney cancer might reconcile the observed immunogenicity of this tumour type despite its low overall mutational burden.

carcinoma; anti-PD-1), carcinoma of the bladder (anti-PD-L1), and head and neck squamous cell carcinoma (anti-PD-1), as well as microsatellite instability high (MSI-H) tumours of any tissue subtype. T cells reactive to tumour-specific mutant antigens (neoantigens) have been detected across the common epithelial malignancies[4] and neoantigens are increasingly shown to be the target of checkpoint inhibitor-induced T-cell responses[5,6] and adoptively transferred T cells.[7–9] Many investigators are leveraging whole-exome sequencing and RNA sequencing, focusing on non-synonymous SNVs (nsSNVs), to predict expressed mutated peptides that bind MHC class I molecules (SNV-neoantigens). Neoantigen burden is closely related to the nsSNV burden, which varies significantly across cancer types, from one nsSNV in paediatric tumours to more than 1500 nsSNVs in tumours associated with microsatellite instability.[10] However, less than 1% of the nsSNVs in expressed genes lead to detectable CD4-positive[11] or CD8-positive T-cell[7] reactivities in tumour-infiltrating lymphocytes. Accordingly, efficacy of checkpoint inhibitors is most marked in tumour types with a high nsSNV burden, including melanoma, lung adenocarcinoma, lung squamous cell carcinoma, head and neck squamous cell carcinoma, and carcinoma of the bladder,[10] which reflects a higher probability of creating a neoantigen that will be presented to and recognised by T cells. Furthermore, within these tumour types, nsSNV and neoantigen burdens correlate with response to checkpoint inhibitors.[12–16] A notable outlier is renal clear cell carcinoma, which has a relatively low nsSNV burden (around ten times lower than melanoma). Renal clear cell carcinoma is characterised by a high level of tumour-infiltrating immune cells[17] and has been shown to respond to interferon-α, high-dose interleukin 2,[18,19] and, more recently, checkpoint inhibitors,[20,21] but the mutational and antigenic determinants of these responses are unknown.

Indel mutations that cause a frameshift (frameshift indels) create a novel open reading frame and could produce a large quantity of neoantigenic peptides highly distinct from self (appendix p 5). It has been hypothesised[22] that novel open reading frames might be an ideal source of tumour-derived neoantigens and so induce multiple neoantigen reactive T cells, because of both an increased number of mutant peptides and reduced susceptibility to self-tolerance mechanisms. On this basis, we aimed to characterise the pattern of indel mutations with pan-cancer analysis and investigate their association with antitumour immune response and outcome following checkpoint blockade.

See Online for appendix

For more on **The Cancer Genome Research Network** see http://cancergenome.nih.gov/

Methods

Study design and participants

Pan-cancer somatic mutational data were obtained from The Cancer Genome Atlas (TCGA) for whole-exome sequencing data of 5777 solid tumours, across 19 cancer types: bladder urothelial carcinoma, invasive breast carcinoma, cervical and endocervical cancers, colorectal adenocarcinoma, glioma, head and neck squamous cell carcinoma, chromophobe renal cell carcinoma, renal clear cell carcinoma, renal papillary cell carcinoma, liver hepatocellular carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, ovarian serous cystadeno-

carcinoma, pancreatic adenocarcinoma, prostate adenocarcinoma, skin cutaneous melanoma, stomach adenocarcinoma, thyroid carcinoma, and uterine carcinosarcoma. We extracted patient-level mutation annotation files from the Broad Institute TCGA GDAC Firehose repository, which had been previously curated by TCGA analysis working group experts to ensure strict quality control. We performed replication analyses in two additional cohorts of patients with renal clear cell carcinoma: a whole-exome sequencing study of 106 patients with renal clear cell carcinomas reported by Sato and colleagues[23] and a whole-exome sequencing study of ten patients with renal clear cell carcinomas reported by Gerlinger and colleagues.[24] We obtained final post-quality control patient-level mutation annotation files for each study.

To further test for an association between nsSNVs or indel loads and patient response to checkpoint inhibitor therapy we used four patient cohorts. The first dataset consisted of 38 patients with melanoma treated with anti-PD-1 therapy, as reported by Hugo and colleagues.[25] We obtained final post-quality control mutation annotation files and clinical outcome data, and 34 patients were retained for analysis after exclusion of cases in which DNA had been extracted from patient-derived cell lines and patients in whom tissue tumor purity was below 20%. Four samples from Hugo and colleagues[25] were taken after a short period on treatment, which raises the possibility that checkpoint inhibitor therapy itself might have affected mutational frequencies through possible elimination of immunogenic tumour clones. To be consistent with the original study, these samples were not excluded; however, we note the frameshift indel association presented becomes more significant with these cases removed. The second checkpoint inhibitor cohort comprised of 62 patients with melanoma treated with anti-CTLA-4 therapy, as reported by Snyder and colleagues.[13] All patients' samples were taken from fresh snap frozen tumour tissue with tumour purity of more than 20% so, accordingly, all 62 cases were retained for analysis. The Snyder and colleagues' cohort also contained a number of samples taken on treatment; these samples have been retained for consistency; however, we note again the significance of results strengthens if they are removed. The third checkpoint inhibitor cohort comprised of 100 patients with melanoma treated with anti-CTLA-4 therapy, as reported by Van Allen and colleagues;[12] one patient (Pat21) was excluded because of a tumour purity of less than 20%. The final checkpoint inhibitor cohort comprised of 31 patients with non-small-cell lung cancer treated with anti-PD-1 therapy, as reported by Rizvi and colleagues;[14] all patients were eligible for inclusion. For these four cohorts,[12–14] final mutation annotation files, including indel mutations, were not available, so we obtained raw BAM files and undertook variant calling using a standardised bioinformatics pipeline. To assess for a general association between nsSNVs or indel loads and patient overall survival we used a final cohort of 100 patients with non-small-cell lung cancer, as reported by Jamal-Hanjani and colleagues.[26] We obtained final post-quality control mutation annotation files and clinical outcome data, and 88 patients were retained for analysis after exclusion of non-smokers. Non-smokers were excluded on account of differing cause of disease and dramatic differences in mutation counts, which were likely to confound analyses. We additionally considered clonal versus subclonal analysis of indel counts in this cohort; however, because of small indel numbers it was not possible to reliably subset the data in this manner.

For more on the **Broad Institute TCGA GDAC Firehose repository** see https://gdac.broadinstitute.org/

Procedures

For whole-exome sequencing variant calling, we obtained BAM files representing both the germline and tumour samples from the cohorts of Snyder and colleagues,[13] Van Allen and colleagues,[12] and Rizvi and colleagues[14] and converted these to FASTQ format using Picard tools (version 1.107) SamToFastq. Raw paired-end reads (100 bp) in FastQ format were aligned to the full hg19 genomic assembly (including unknown contigs) obtained from GATK bundle (version 2.8),[27] using bwa mem (bwa-0.7.7).[28] We used Picard tools to clean, sort, and merge files from the same patient sample and to remove duplicate reads. We used Picard tools, GATK (version 2.8.1), and FastQC (version 0.10.1) to produce quality control metrics. SAMtools mpileup (version 0.1.19)[29] was used to locate non-reference positions in tumour and germline samples. Bases with a Phred score of less than 20 or reads with a mapping quality less than 20 were omitted. Base-alignment quality computation was disabled and the coefficient for downgrading mapping quality was set to 50. VarScan2 somatic (version 2.3.6)[30] used output from SAMtools mpileup to identify somatic variants between tumour and matched germline samples. Default parameters were used with the exception of minimum coverage for the germline sample, which was set to 10, and minimum variant frequency was changed to 0·01. VarScan2 processSomatic was used to extract the somatic variants. The resulting SNV calls were filtered for false positives with the associated fpfilter.pl script in Varscan2, initially with default settings then repeated with min-var-frac=0·02, having first run the data through bam-readcount (version 0.5.1). Only indel calls classed as high confidence by VarScan2 processSomatic were kept for further analysis, with somatic_p_value scores less than 5×10^{-4}. MuTect (version 1.1.4)[31] was also used to detect SNVs, with annotation files contained in GATK bundle. Following completion, variants called by MuTect were filtered according to the filter parameter PASS.

In the pan-cancer cohort, SNV and indel mutation counts were computed per case, considering all variant types. Across all 5777 samples, we observed a total of

For more on **Picard tools** see http://broadinstitute.github.io/picard

For more on **FastQC** see http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

For more on **bam-readcount** see https://github.com/genome/bam-readcount

1 227 075 SNVs and 54 207 indels. Dinucleotide and trinucleotide substitutions were not considered. The metric indel burden was simply defined as the absolute indel count per case and indel proportion was defined as follows

$$\text{indel proportion} = \frac{\text{number of indels}}{(\text{number of indels} + \text{number of SNVs})}$$

We repeated the same analysis in the two renal clear cell carcinoma replication cohorts.

We estimated non-sense-mediated decay (NMD) efficiency with RNAseq expression data (as measured in transcripts per kilobase million), obtained from the TCGA GDAC Firehose repository. We estimated the extent of NMD for all indel and SNV mutations (with SNV mutations used as a benchmark comparator) by comparing mRNA expression in samples with a mutation to the median mRNA expression of the same transcript across all other tumour samples in which the mutation was absent. Specifically, mRNA expression of every mutation-bearing transcript was divided by the median mRNA expression of that transcript in non-mutated samples, to give an NMD index. The overall NMD index values observed were 0·93 (indels) and 1·00 (SNVs), suggesting an overall 7% reduction in expression in indel mutated transcripts. Tumour purity in the renal clear cell carcinoma cohort was 0·54,[32] quantified by histological assessment, and assuming constant expression in the remaining 0·46 normal cellular content, that would yield an adjusted 14% drop in expression in indel-mutation-bearing cancer cells. If we assume that tumour mutations are clonal, of heterozygote genotype, in a diploid genomic region, and wild-type allele expression in mutated cancer cells remains constant, a purity-adjusted reduction of 0·5 would be expected under a model of fully effective NMD. Hence these data suggest NMD operates with reduced efficiency in the renal clear cell carcinoma cohort; however, we acknowledge that these assumptions will have some effect. These data are presented as a global approximation of NMD efficiency, using methods in line with previous publications.[33] NMD index values were $-\log_2$ transformed, with 0 indicating no mRNA degradation and plotted for indel or SNV mutations.

We used PyClone[34] and ASCAT[35] to determine the clonal status of variants in the cohorts by Snyder and colleagues[13] and Van Allen and colleagues.[12] For each case variant calls were integrated with local allele-specific copy number (obtained from ASCAT), tumour purity (also obtained from ASCAT), and variant allele frequency. All mutations were then clustered using the PyClone Dirichlet process clustering. We ran PyClone with 10 000 iterations and a burn-in of 1000, and default parameters. For a number of tumours the reliable copy number, mutation, and purity estimations could not be extracted, rendering clonal architecture analysis intractable and these tumours were omitted from the analysis. The following sample was excluded because of an absence of accurate copy number or clonality estimation in Snyder and colleagues' cohort[13]: V_MSK052. For Van Allen and colleagues' cohort,[12] reliable analysis of indel mutation clonality was not possible because of a lack of accurate copy number or clonality estimation in a number of cases: Pat02, Pat06, Pat100, Pat101, Pat103, Pat106, Pat110, Pat113, Pat131, Pat132, Pat135, Pat138, Pat139, Pat140, Pat148, Pat159, Pat160, Pat163, Pat165, Pat166, Pat170, Pat171, Pat174, Pat175, Pat24, Pat36, Pat38, Pat73, Pat77, Pat78, Pat79, and Pat92.

For a subset of patients (n=4592) from the TCGA cohort, tumour-specific neoantigen binding affinity prediction data were also available and obtained from Rooney and colleagues.[36] Briefly, the four digit HLA type for each sample, along with mutations in class I *HLA* genes, were determined using POLYSOLVER (POLYmorphic loci reSOLVER).[37] We determined somatic mutations using Mutect[31] and Strelka[38] tools. All possible 9-mer and 10-mer mutant peptides were computed, on the basis of the detected somatic SNV and indel mutation across the cohort. Binding affinities of mutant and corresponding wild-type peptides, relevant to the corresponding POLYSOLVER-inferred HLA alleles, were predicted using NetMHCpan (version 2.4).[39] High-affinity binders were defined as IC_{50} less than 50 nM. Wild-type allele non-strong binding was defined as IC_{50} greater than 50 nM. Accordingly a mutant-specific binder was used to refer to a neoantigen with mutant IC_{50} less than 50 nM and wild-type IC_{50} more than 50 nM. A strong binding threshold was used for wild-type alleles to ensure fair comparison between SNV-derived and indel-derived neoantigens, in view of the high incidence of wild-type non-binders for indels. We excluded (from the pan-cancer neoantigen analyses) cancers that were associated with a high level of viral genome integration, including cervical (>80% rate of human papillomavirus integration) and hepatocellular carcinoma (>50% rate of hepatitis B integration), but not head and neck squamous cell carcinoma (<15% rate of human papillomavirus integration). No TCGA dataset was available for Merkel cell carcinoma.

Immune gene signature data were obtained from Rooney and colleagues,[40] with gene sets defined as stated in the appendix (p 3). We did analysis for TCGA patients with renal clear cell carcinoma (n=392), for whom both RNAseq and neoantigen data were available. A high burden of frameshift indel high-affinity neoantigens was defined as more than 10 per case (n=32), and the percentage difference in expression was compared between the high indel neoantigen group and all other patients across each immune signature. We excluded immune signatures with minimal ssGSEA enrichment scores (<0·5) in all groups. The same analysis was repeated for a high burden of SNV-derived high-affinity

neoantigens, with a threshold of more than 17 SNV neoantigens selected to size match the high burden groups (equal number of patients; n=32 across all high-load groups) across mutational types. We plotted the percentage differences in expression in heatmap format. We did correlation analysis within the high-frameshift indel neoantigen group (n=32 patients with renal clear cell carcinoma).

Outcomes

Across the four cohorts of patients treated with checkpoint inhibitors, we tested nsSNV, all-coding indel, and frameshift indel variant counts for an association with patient response to therapy. For each of these measures, high groups were defined as the top quartile and low groups were defined as the bottom-three quartiles. We used the same criteria across all four datasets and compared the proportion of patients responding to therapy in high and low groups. Measures of patient response were based on definitions consistent with how they were evaluated in the said trials, as follows. For Snyder and colleagues' cohort,[13] long-term clinical benefit was defined as radiographic evidence of freedom from disease, evidence of a stable disease, or decreased volume of disease for more than 6 months. No long-term clinical benefit was defined as tumour growth on every CT scan after the initiation of treatment (no benefit) or a clinical benefit lasting 6 months or less (minimal benefit). For Hugo and colleagues' cohort,[25] responding tumours were complete response, partial response, and stable disease, and non-responding tumours were defined as disease progression. For Van Allen and colleagues' cohort,[12] clinical benefit was defined as complete response, partial response, or stable disease, and no clinical benefit was progressive disease or stable disease with overall survival less than 1 year. For Rizvi and colleagues' cohort,[14] durable clinical benefit was defined as partial response or stable disease lasting longer than 6 months, and no durable benefit was progressive disease less than 6 months from beginning of therapy.

Statistical analysis

Survival analysis was done using the Kaplan-Meier method, with p value determined by a log-rank test. Relapse-free survival was defined as the time to recurrence or relapse, or if a patient had died without recurrence, the time to death. Hazard ratio (HR) was determined through a Cox proportional hazards model. Multivariate Cox regression was done with relapse-free survival versus indel load with stage, adjuvant therapy (yes or no), age, and histology included in the model.

We compared indel burden and proportion measures between renal cell carcinomas and all other non-kidney cancers with a two-sided Mann-Whitney *U* test. In the checkpoint inhibitor response analysis, nsSNV, exonic indel, and frameshift indel counts were each compared to patient response outcome using a two-sided Mann-Whitney *U* test. We did a meta-analysis of results across the four checkpoint inhibitor datasets using the Fisher's method of combining p values from independent tests. We undertook immune signature correlation analysis using a Spearman's rank correlation coefficient. We carried out statistical analyses using R (version 3.0.2) and considered a p value of 0·05 or less (two-sided) as being statistically significant.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

We observed a median indel proportion value of 0·05 and a median indel count of 4, cohort-wide. Across all tumour types, renal clear cell carcinoma was found to have the highest proportion of coding indels, 0·12 ($p<2·2\times10^{-16}$; figure 1), a 2·4 times increase when compared with the pan-cancer average. This result was replicated in two further independent cohorts, with median observed indel proportions of 0·10 in Sato and colleagues' study[23] and 0·12 in Gerlinger and colleagues' study[24] (figure 1). Renal papillary cell carcinoma and chromophobe renal cell carcinoma had the second and third highest indel proportion, suggesting a possible tissue-specific mutational process contributing to the acquisition of indels in renal cancers. Renal papillary cell carcinoma (median indel number of 10 [95% CI 9–11]) and chromophobe renal cell carcinoma (8 [7–10]) had the highest absolute indel count across all tumour types, closely followed by renal clear cell carcinoma (7 [6–8]). Renal clear cell carcinoma is characterised by loss-of-function mutations in one or more tumour-suppressor genes: *VHL, PBRM1, SETD2, BAP1,* and *KDM5C*,[32] which can be inactivated by nsSNV or indel mutations. To exclude the possibility that these hallmark mutations were distorting the results, we recalculated renal clear cell carcinoma indel proportion excluding these genes; the revised indel proportion remained at 0·12. When we used previously published multiregion whole-exome sequencing data[24] from ten cases of renal clear cell carcinoma to assess the clonal nature of indel mutations, 53 (48%) of 110 frameshifting indels were clonal in nature (present in all tumour regions).

The overall effect of NMD on the expression of indel-mutated genes was estimated to be 14% (7% drop divided by 0·54 tumour purity), suggesting it operates on a subset of transcripts (appendix p 6).

Next we sought to investigate the potential immunogenicity of nsSNV and indel mutations through analysis of MHC class I-associated tumour-specific neoantigen binding predictions in the pan-cancer TCGA cohort.

Across all samples, HLA-specific neoantigen predictions were done on 335 594 nsSNV mutations, resulting in a total of 214 882 high-affinity binders (defined as epitopes with predicted IC_{50} <50 nM; the concentration necessary to reduce the binding affinity by half), equating to a rate of 0·64 neoantigens per nsSNV mutation (SNV-neoantigens; table). In a similar manner, predictions were made on 19 849 frameshift indel mutations, resulting in 39 768 high-affinity binders with a rate of 2·00 neoantigens per frameshift mutation (frameshift neoantigens; table). Thus on a per mutation basis, frameshift indels could generate around three times more high-affinity neoantigen binders than nsSNVs (table), consistent with the prediction in a recent analysis of a colorectal cancer cohort.[41] When both wild-type and mutant peptides are predicted to bind, central immune tolerance mechanisms might delete cells with the reactive T-cell receptor.[42] Therefore, we repeated a pan-cancer analysis restricting the neoantigens to mutant-specific binders (ie, where the wild-type peptide is not predicted to be a strong binder), and showed that frameshift indels were nine times enriched for mutant-allele-only binders (table).

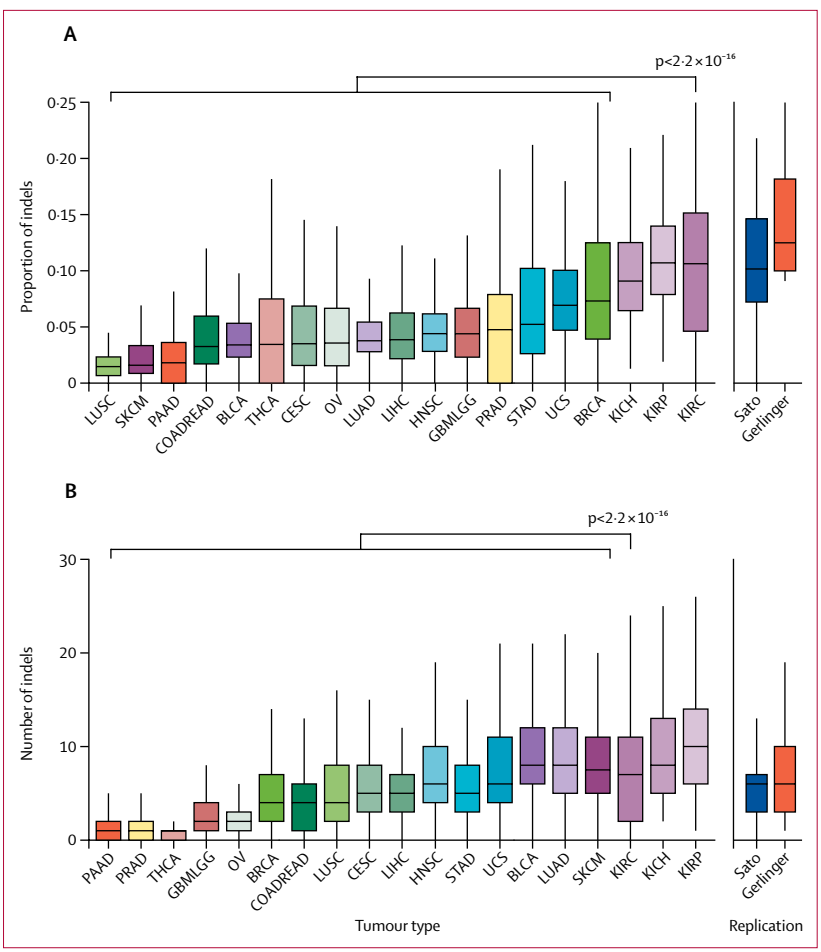


Figure 1: **Occurrence of indels in different types of solid tumour**
Proportion (A) and absolute count (B) of indel mutations across 19 solid tumour types from TCGA. The last two boxplots are additional independent renal clear cell carcinoma replication datasets from Sato and colleagues[23] and Gerlinger and colleagues.[24] Statistical comparison is renal clear cell carcinoma cohort compared with all other non-kidney TCGA samples. indel=insertions and deletions. TCGA=The Cancer Genome Atlas. LUSC=lung squamous cell carcinoma. SKCM=skin cutaneous melanoma. PAAD=pancreatic adenocarcinoma. COADREAD=colorectal adenocarcinoma. BLCA=bladder urothelial carcinoma. THCA=thyroid carcinoma. CESC=cervical and endocervical cancers. OV=ovarian serous cystadenocarcinoma. LUAD=lung adenocarcinoma. LIHC=liver hepatocellular carcinoma. HNSC=head and neck squamous cell carcinoma. GMBLGG=glioblastoma multiforme and low-grade glioma. PRAD=prostate adenocarcinoma. STAD=stomach adenocarcinoma. UCS=uterine carcinosarcoma. BRCA=invasive breast carcinoma. KICH=chromophobe renal cell carcinoma. KIRP=renal papillary cell carcinoma. KIRC=renal clear cell carcinoma.

	Mutations (n)	Neoantigens (n)*	Mutant-specific neoantigens (n)†	Neoantigens per mutation	Mutant-specific neoantigens per mutation
nsSNVs	335 594	214 882	75 224	0·64	0·22
fs-indels	19 849	39 768	39 608	2·00	2·00
Enrichment	..	..	..	3·13	8·94

nsSNVs=non-synonymous single nucleotide variants. fs-indels=frameshift insertions and deletions. *Strong binders (<50 nM affinity). †Wild-type allele non-strong binding (>50 nM affinity).

Table: **Neoantigens per variant class**

Of particular interest were genes that are frequently altered via frameshift mutations and with high propensity for MHC binding. In a pan-cancer analysis, these genes were enriched for classic tumour-suppressor genes, including *TP53*, *ARID1A*, *PTEN*, *KMT2D*, *KMT2C*, *APC*, and *VHL* (figure 2). Collectively, the top 15 genes with the highest number of frameshift mutations were mutated in more than 500 samples (approximately 10% of the cohort with 5777 samples) with more than 2400 high-affinity neoantigens predicted. Tumour-suppressor genes have been a previously intractable mutational target, but they might be targetable as potent neoantigens. Furthermore, by virtue of being founder events, many alterations in tumour-suppressor genes are clonal, present in all cancer cells, rendering them compelling targets for immunotherapy.[43]

We next considered the clinical effect of indel mutations by assessing the association between neoantigen enrichment and therapeutic benefit. Consistent with a potential role of frameshifts in the generation of neoantigens, those tumour types approved for the use of checkpoint inhibitors were all found to harbour an above average number of frameshift neoantigens, despite substantial differences in the total SNV or indel mutational burden—eg, renal cell carcinoma (figure 3). Overall, the number of frameshift neoantigens were significantly higher in the checkpoint inhibitor-approved tumour types versus those that have not been approved to date ($p<2·2\times10^{-16}$). However, the potential presence of frameshift neoantigens alone does not imply that they induce T-cell responses, and hence we tested their effect on checkpoint inhibitor efficacy. We used the exome sequencing results from an anti-PD-1 study[25] in melanoma (n=38 patients). We tested three classes of mutation, nsSNVs, in-frame indels, and frameshift indels, for an association with response to treatment. Although nsSNVs (p=0·27) and in-frame (3n) indels

(p=0·19) had no association with response to treatment, frameshift indel mutations were significantly associated with anti-PD-1 response (p=0·023; figure 4A). The upper quartile of patients with the highest burden of frameshift indels had an 88% (seven of eight cases) response to anti-PD-1 therapy, compared with 43% (11 of 26 cases) for the lower three quartiles (odds ratio 9·5 [95% CI 1·02–89·23]; figure 4B). To confirm the reproducibility of this association, further checkpoint inhibitor response data were obtained from two additional melanoma cohorts: Snyder and colleagues' cohort[13] (n=62, anti-CTLA-4 treated) and Van Allen and colleagues' cohort[12] (n=100, anti-CTLA-4 treated). We did the same analysis in each cohort and frameshift indel burden was significantly associated with checkpoint inhibitor response in both datasets (HR 3·4 [95% CI 1·05–11·27]; p=0·0074 for Snyder and colleagues' cohort and 2·9 [1·15–7·55]; p=0·032 for Van Allen and colleagues' cohort; figure 4A). An overall meta-analysis across the three cohorts confirmed frameshift indel count to be associated with checkpoint inhibitor response ($p=4·7\times10^{-4}$), and with a more significant association than nsSNV count ($p=4·8\times10^{-3}$). The effect of clonality was additionally assessed, and clonal frameshift indels were found to have a further significantly predictive advantage beyond all frameshift indels (clonal and subclonal; appendix p 7), supporting previous work reported by our group.[43] Overall survival analysis was not different between high and low frameshift indel groups, possibly because of the effect of subsequent therapies on the overall survival (OR 2·43 [95% CI 0·77–7·73]; p=0·228 for Hugo and colleagues' cohort[25]; appendix p 8). We assessed the association between frameshift indel load and checkpoint inhibitor response in another tumour type by using data obtained from Rizvi and colleagues' small cohort[14] of 31 patients with non-small-cell lung cancer treated with anti-PD-1 therapy; no difference was observed (p=0·23). To further investigate the importance of frameshift indels in non-small-cell lung cancer, we did additional analysis using data from Jamal-Hanjani and colleagues' cohort[26] of 100 cases, none of whom received treatment with checkpoint inhibitors. Consistent with our previous findings,[43] we observed that patients with lung adenocarcinoma whose tumour's harboured a high clonal neoantigen burden (higher than upper quartile of cohort) exhibited improved relapse-free survival compared with the bottom three quartiles (p=0·026). However, across all histological subtypes of non-small-cell lung cancer, survival was found to be significantly improved for patients with a high load of frameshift indels (*vs* low load: HR 0·25 [95% CI 0·06–1·08]; p=0·045); by contrast, nsSNV load was not formally associated (0·36 [0·11–1·21]; p=0·084; appendix p 9). Of note, the strongest prognostic predictor was for patients in the patients with a high load of both nsSNVs and frameshift indels, with elevated levels of both frameshift indels and nsSNVs, with no events in this group

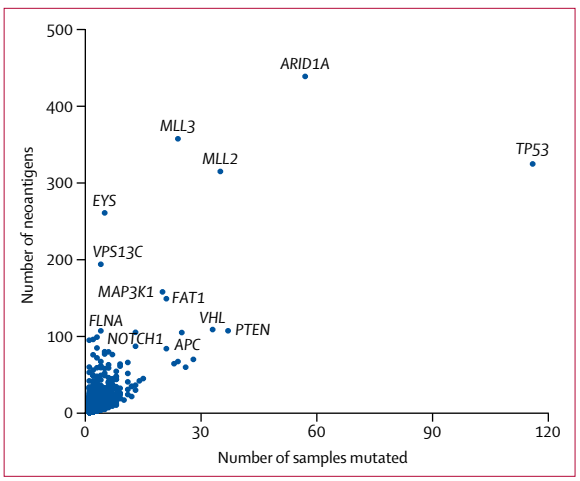
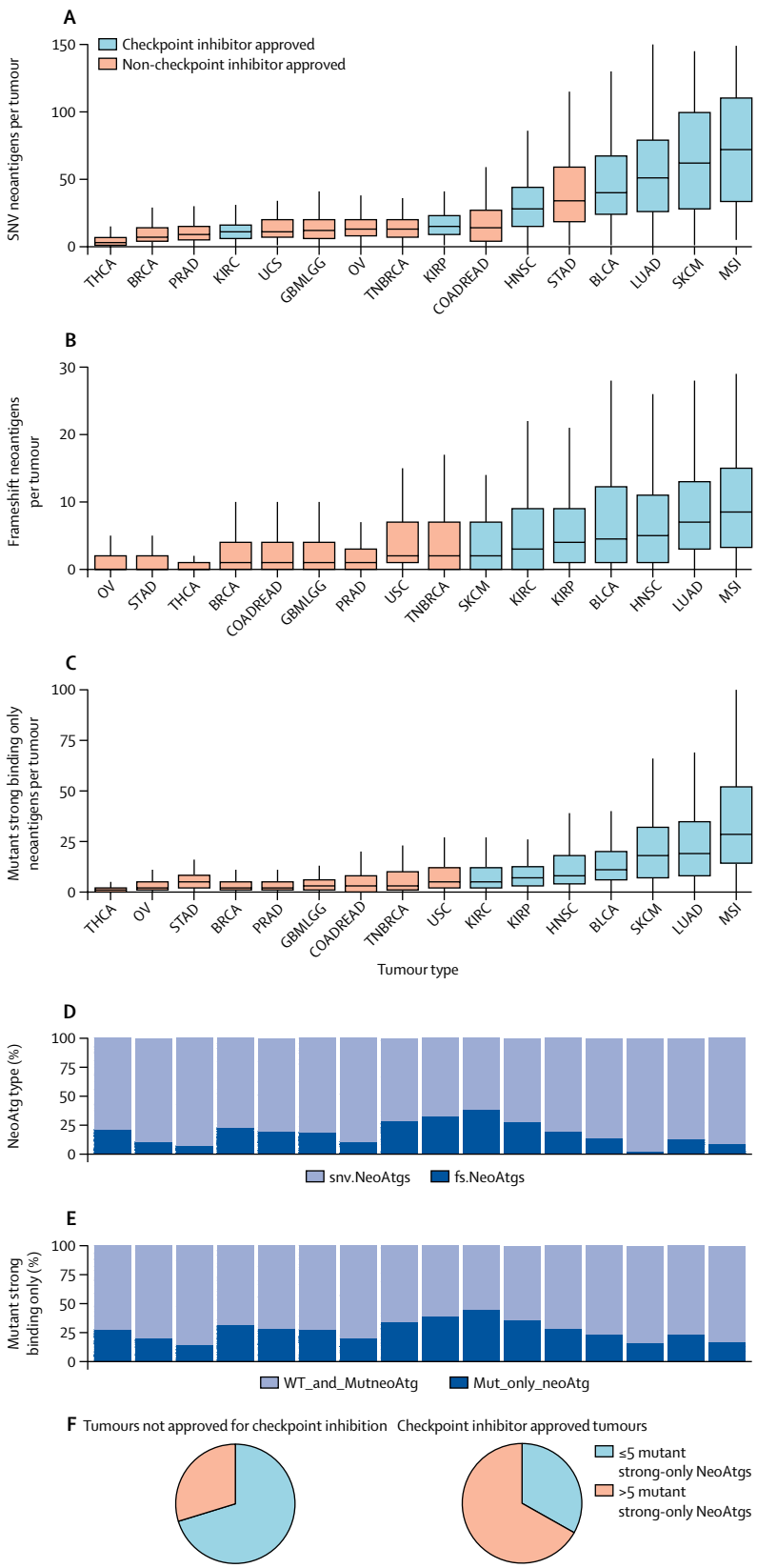


Figure 2: **Recurrent genes with frameshift indel neoantigens**
Data are from all patients in The Cancer Genome Atlas pan-cancer cohort. The graph shows the number of unique samples containing a frameshift indel neoantigen and the number of unique neoantigens (ie, each mutation can generate multiple neoantigens). Marked are genes either mutated in more than 30 samples or with more than 80 neoantigens.

(p=0·025). Multivariate analysis showed some evidence of correlation between variables (appendix p 2), so further investigation of nsSNVs and frameshift indels as predictors in larger patient cohort will be required to draw definitive conclusions.

Analyses of the indel load and proportion of response achieved from phase 2 studies for the tumour types not approved for checkpoint inhibition were limited by the small sample size and variable patient inclusion criteria such as PDL-1 immunohistochemistry (appendix p 4). Nevertheless, the proportion of patients achieving a response was higher in triple-negative breast cancer[44] compared with other invasive breast carcinoma molecular subtypes, and triple-negative invasive breast carcinoma has a higher burden of frameshift and mutant-specific neoantigens (figure 3). Furthermore, mutational burden has been reported as higher in *BRCA1*-mutated triple-negative breast cancer compared with *BRCA*-wild-type triple-negative breast cancer,[45] and we specifically observed a higher indel load in these cases (appendix p 10). However, this outcome did not correlate with tumour-infiltrating lymphocyte density (appendix p 10), possibly because of the small sample size, absence of indel immunogenicity in this tissue type, or additional factors that modulate tumour-infiltrating lymphocyte density.

Finally, although genomic data are not available to correlate with checkpoint inhibitor response in renal clear cell carcinoma, we analysed the association between frameshift neoantigen load and immune responses within the tumour using RNAseq gene expression data. Patients were split into groups on the basis of the burden of frameshift neoantigens (high defined as >10 frameshifts per case, with this threshold set to capture the top 10% of cases) versus SNV-neoantigens



(high defined as >17 nsSNVs per case, with this threshold set to ensure matched patient sample sizes). A high load of frameshift neoantigens was associated with upregulation of immune signatures classically linked to immune activation, including MHC class I antigen presentation, CD8-positive T-cell activation, and increased cytolytic activity, a pattern not observed in the high SNV-neoantigen group (figure 5). Furthermore, correlation analysis within the high frameshift neoantigen group showed that CD8-positive T-cell signature was correlated with both MHC class I antigen presentation genes (r=0·78) and cytolytic activity (r=0·83; figure 5).

Discussion

In this study, we analysed the pattern of indel mutations across 19 solid tumour types and found that renal clear cell carcinoma, renal papillary cell carcinoma, and chromophobe renal cell carcinoma have the highest indel rate as a proportion of their total mutational burden and the highest overall indel count and are enriched for mutant-specific neoantigens. We also observed that indel number is significantly associated with checkpoint inhibitor response in melanoma.

Indels are thought to occur as a result of DNA strand slippage during DNA synthesis[46] and their frequency is higher in repetitive sequences, especially those that are AT-rich. Indels are also generated through mutagen exposure, with a higher number observed in smoking than in non-smoking non-small-cell lung cancer (lung adenocarcinoma)[40] and higher in UV-exposed (cutaneous) versus UV-protected (mucosal) melanomas.[47] Less is known about the repair of indels than SNVs; however, the role of the mismatch repair mechanism is illustrated by the microsatellite instability-high phenotype, characterised by excess indels in repetitive sequences as seen in patients with Lynch syndrome. Although renal clear cell carcinoma has been reported in patients with Lynch syndrome,[48] this cannot account for the overall pattern of indel rates across renal clear cell

Figure 3: **Tumour-specific neoantigen counts by cancer type**
Count of SNV-derived neoantigens (A), frameshift indel-derived neoantigens (B), and mutant-only neoantigen binders (C), and proportion of neoantigens derived from SNVs and indels (D) and neoantigens in which mutant allele-only binds (E). (F) Proportion of samples with more than or less than five mutant-only neoantigen binders. The threshold was set at five because all checkpoint approved cancer types had median values above this level. (A–C) Data are ordered by median value, from lowest to highest. (D, E) Data are ordered the same as (C). SNV=single nucleotide variant. indel=insertions and deletions. THCA=thyroid carcinoma. BRCA=invasive breast carcinoma. PRAD=prostate adenocarcinoma. KIRC=renal clear cell carcinoma. UCS=uterine carcinosarcoma. GMBLGG=glioblastoma multiforme and low-grade glioma. OV=ovarian serous cystadenocarcinoma. TNBRCA=triple-negative invasive breast carcinoma. KIRP=renal papillary cell carcinoma. COADREAD=colorectal adenocarcinoma. HNSC=head and neck squamous cell carcinoma. STAD=stomach adenocarcinoma. BLCA=bladder urothelial carcinoma. LUAD=lung adenocarcinoma. SKCM=skin cutaneous melanoma. MSI=microsatellite instability. NeoAtgs=neoantigens. fs=frameshift. WT=wild type. Mut=mutant.

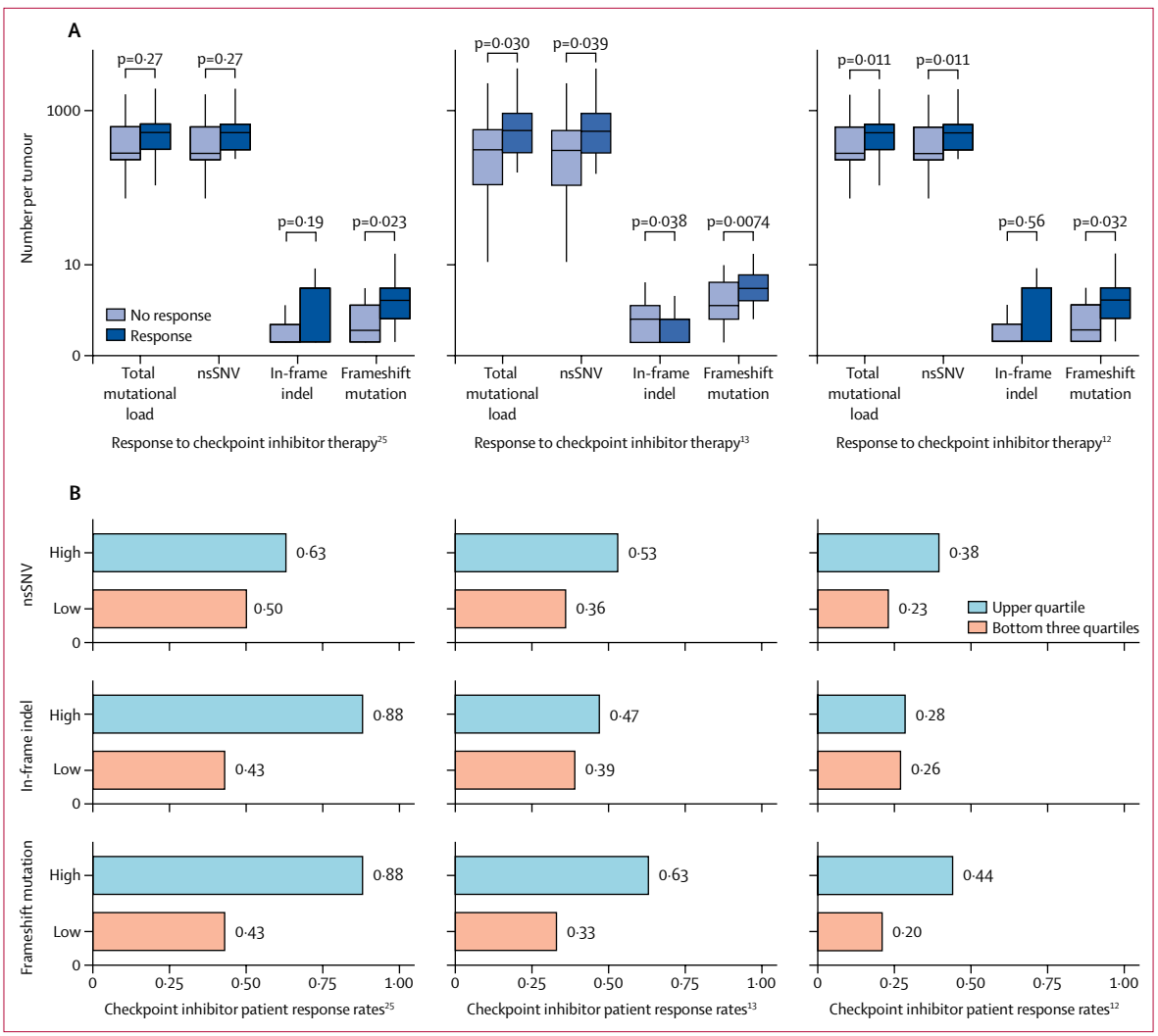


Figure 4: **Checkpoint inhibitor patient response**
Patients were separated by their response to checkpoint inhibitor therapy (A) and split into high and low groups as per the y-axis-labelled metrics (B) in the three melanoma cohorts.[12,13,25] nsSNV=non-synonymous single nucleotide variant.

carcinoma nor the comparatively low SNV burden. Most renal clear cell carcinomas have loss of chromosome 3p, which encodes the mismatch repair gene *MLH1*, but the remaining allele is rarely mutated in sporadic renal clear cell carcinoma. Another relevant gene encoded on 3p is *FHIT*, and its deficiency has been linked with indel accumulation in knockout mouse models, but the consequences of the heterozygous knockout (whether haploinsufficient) are unknown.[49] However, as loss of 3p is an infrequent event in renal papillary cell carcinoma and chromophobe renal cell carcinoma and indels are also elevated in both these tumour types, other tissue-specific phenomena are likely to contribute to the increased indel burden across all renal carcinoma subtypes.[50] Renal clear cell carcinoma and renal papillary cell carcinoma arise in the proximal tubule and chromophobe renal cell carcinoma in the distal tubule of the nephron, and this shared tissue context might be important, even if the three subtypes are molecularly distinct.[32,50,51] The nephron, and the proximal tubule in particular, play a crucial role in the reabsorption of vast volumes of renal filtrate and elimination of waste products of metabolism and toxins, with the effects of toxin elimination evident in the increased incidence of renal clear cell carcinoma in those individuals exposed to aristolochic acid.[52] Ochratoxin A, a mycotoxin, induces renal tumours in rodents by causing double-strand breaks.[53] Polymorphisms in genes involved in the repair of double-strand breaks are associated with an increased risk of renal clear cell carcinoma.[54] Double-strand breaks are mostly repaired by non-homologous end-joining, which is error-prone and can increase the rate of small indels (1–10 bp). Therefore, it is possible that an environmental toxin causes an excess of double-strand

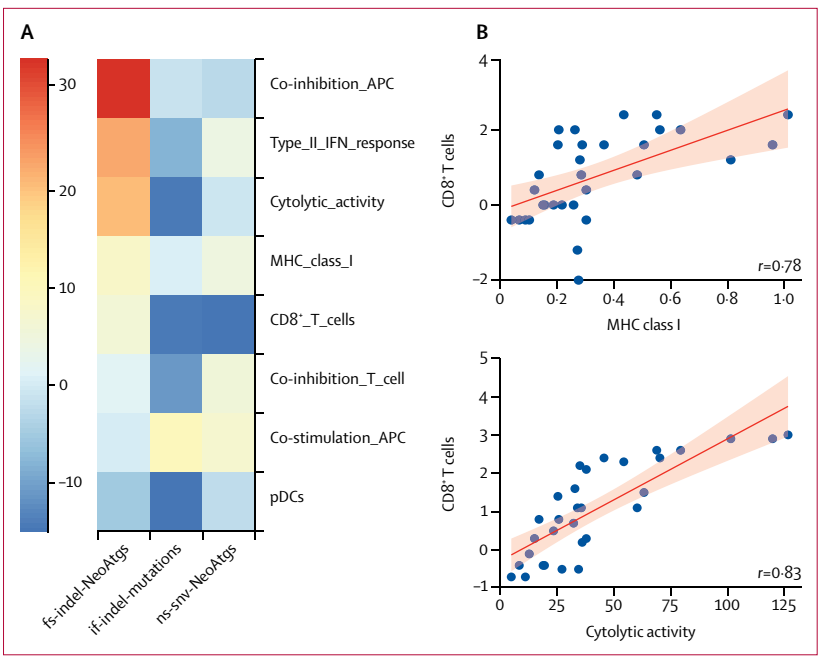


Figure 5: **Immune gene signatures in patients with renal clear cell carcinoma**
(A) Percentage change in median gene signature expression is shown between high and low groups for each metric on the x-axis as labelled. Several pathways were found to be exclusively upregulated in the high fs-indel-NeoAtg group. (B) Correlation analysis within the high fs-indel-NeoAtgs group showed the CD8+ T-cell signature was correlated with both MHC class I antigen presentation genes and cytolytic activity. Shown is the line of best fit and shaded area to represent 95% CI. fs-indel-NeoAtgs=frameshift indel neoantigen count. if-indel-mutations=in-frame indel mutation count. ns-snv-NeoAtgs=non-synonymous single nucleotide variant neoantigen count. APC=adenomatous polyposis coli. IFN=interferon. CD8+=CD8-positive. pDCs=plasmacytoid dendritic cells.

breaks in the nephron and that non-homologous end-joining's mutagenic potential is exacerbated by functional polymorphisms. In support of this notion we observed a higher rate of indels in triple-negative breast cancer, which is enriched for *BRCA* deficiency. BRCA1 has been shown to inhibit error prone non-homologous end-joining.[55] However, we did not observe a correlation between indel load and tumour-infiltrating leucocytes density in *BRCA1* triple-negative invasive breast carcinoma.

We observed that indels, which alter the reading frame, generate three times as many predicted neoantigens as nsSNVs and nine times as many strong mutant-binding neoantigens where the wild-type sequence is not predicted to strongly bind the HLA molecule (IC_{50} >50 nM). Thus, frameshift mutations potentially result in a neoantigen landscape, which is both quantitatively and qualitatively more potent than that provided by an equivalent number of nsSNVs. In keeping with this notion, microsatellite instability-high colorectal cancer CD8-positive tumour-infiltrating leucocytes density correlates positively with the total number of frameshift mutations.[56] With the exception of polyomavirus-positive Merkel cell carcinoma and Hodgkin's lymphoma, renal clear cell carcinoma is the only tumour type with a relatively low nsSNV burden among the tumour types for which checkpoint inhibitors have been approved for clinical use. However, owing to a comparable frameshift burden its level of mutant-specific high-affinity neoantigens is similar to that observed in non-small-cell lung cancer and melanoma, and the same is true of renal papillary cell carcinoma and chromophobe renal cell carcinoma. Although the evidence for the immunogenicity of renal papillary cell carcinoma is sparse, complete responses have been noted with the use of both high-dose interleukin 2[57] and anti-PD-1 therapy.[58,59] Therapeutic data in chromophobe renal cell carcinoma are limited.

Given the differential benefit across patients, the spectrum of immune-related adverse events, and the cost of checkpoint inhibitor drugs, efforts to identify biomarkers of response are ongoing. PDL-1 expression and MSI-H status are the only biomarkers that have been linked to drug approval. Mutational and neoantigen burdens have been shown to correlate with clinical outcomes from checkpoint inhibitor therapy in patients with advanced melanoma, colorectal cancer, and non-small-cell lung cancer.[13,14,16] However, some patients with cutaneous melanoma with a low nsSNV burden still derive benefit from checkpoint inhibitors, as do some patients with UV-protected mucosal melanomas,[60] which have a characteristically low nsSNV burden.[61] We analysed three melanoma datasets for which both response and mutational data were available. In two of the three studies,[13,14] comprising a total of 96 patients treated with either anti-PD-1 or anti-CTLA-4 therapy, frameshift indel burden was a better predictor of response than nsSNV burden. In the third study[12] of 100 patients treated with anti-CTLA-4 therapy, the nsSNV burden and frameshift burden were both significantly associated with checkpoint inhibitor response. We note that most of the patients in Van Allen and colleagues' cohort[12] were pretreated and therefore any mutational biomarker assessment in this group might be less reliable.

Although nsSNVs contribute greatly towards tumour immunogenicity in heavily mutated tumours, our analyses suggest that frameshift mutations also make a significant contribution relative to their overall low number. The contribution of frameshift indels in low nsSNV burden tumours might be of greater importance still, as illustrated by the fact that frameshift mutations contribute over a third of the neoantigen load in renal clear cell carcinoma. Mutational and checkpoint inhibitor response data were not available for renal clear cell carcinoma, hence we could not establish a direct association between frameshift indels and positive checkpoint inhibitor response. In terms of indirect evidence in renal clear cell carcinoma, we observed an association between the frameshift neoantigens and upregulation of machinery necessary for antigen presentation by the MHC complex and T-cell activation. Furthermore, the CD8-positive T-cell signature in the frameshift neoantigen-high group was closely related to cytolytic activity, suggesting the presence of antitumour

effectors that could confer sensitivity to immunotherapy. However, no definitive conclusions can be drawn until checkpoint inhibitor response and indel load is directly investigated in a sufficiently powered series of renal clear cell carcinoma cases.

For frameshift neoantigens to contribute to antitumour immunity the mutant peptides must be expressed. Frameshifts cause premature termination codons and the resultant mRNAs can be targeted for NMD. Published analyses[62] of germline samples show that premature termination codons frequently lead to the loss of expression of the variant allele, but that some mutant transcripts escape NMD on the basis of the exact location of the frameshift within a gene. Combined analyses of mutational and expression data from more than 10 000 cancer samples showed that NMD is triggered with variable efficacy, and even when effective might not alter expression because of factors such as short mRNA half-life.[33] RNAseq analysis in renal clear cell carcinoma cases showed a minimal change in mRNA transcript levels for frameshift indel-mutated tumours, suggesting NMD is operating on a subset of transcripts, as expected. In this context, the strongly hypoxic microenvironment that characterises renal clear cell carcinomas might be a contributing factor, with evidence showing NMD inhibition in cells subject to hypoxia and other perturbed microenvironmental conditions.[63]

Clonal frameshift mutations could be an important source of tumour-specific antigens for personalised immunotherapy strategies, including peptide vaccines and adoptive cell therapy. Tumour-reactive T cells recognising a frameshifted product of the *CDKN2A* tumour-suppressor gene were reported to mediate a potent in-vivo response in melanoma.[64] In microsatellite instability-high colorectal cancer, frameshift neopeptide-specific cytotoxic T-cell responses were observed in patients harbouring those mutations.[65] Cytotoxic T-lymphocyte responses to frameshifted proteins have been detected in healthy hereditary non-polyposis colorectal cancer-mutation carriers, raising the possibility of protective immunosurveillance in this population.[66] Frameshift neoantigens are particularly pertinent in the context of mismatch repair-deficiency, which is a pan-cancer event, and crucially, frameshifts commonly occurring in microsatellite instability-high colorectal carcinomas have been shown to generate NMD-resistant transcripts.[67] In support of this, in a study[68] of PD-1 blockade in patients with microsatellite instability-high tumours from various cancer subtypes, functional analyses in a responding patient showed in-vivo expansion of frameshift neoantigen-specific T-cell clones.

Frameshift neoantigens provide a unique opportunity to target common tumour-suppressor genes, such as such as *TP53* and *BAP1*,[69] and their founder status also enriches for clonal neoantigens. Acknowledging the qualitative difference in the neoantigen burden of renal clear cell carcinoma might be integral for optimising responses to checkpoint inhibitors. Neoantigens derived from driver mutations elicit profound T-cell exhaustion via chronic antigen stimulation, generating T-cell pools refractory to immune therapy.[70] Thus, early administration of checkpoint blockade might further improve clinical benefit in cancers with particularly antigenic mutations such as renal clear cell carcinoma. It is also noteworthy that a high differential affinity between wild-type and mutant peptides is indicative of enhanced tumour protection in vivo.[71] The enrichment of mutant-only binders by nine times in neoantigens derived from frameshift mutations relative to nsSNVs might therefore partly explain the predictive power of frameshift neoantigens in checkpoint inhibitor responses.

A widely recognised challenge in bioinformatics is indel variant calling, due to the inherent nature of short-read sequencing technology; however, accurate indel calling can be achieved within both a research and clinical context with strict quality control procedures.[72] While strict quality control procedures can ensure a low false-positive rate, as a consequence the true rate of indel mutations might be underestimated.

In conclusion, we report that kidney cancers carry the highest pan-cancer burden of indel mutations. Futhermore, our data suggest that frameshift indels are a highly immunogenic mutational class; triggering an increased quantity of neoantigens and greater mutant binding specificity. Collectively, these data might reconcile the outlier nature of immunotherapy responses in renal clear cell carcinoma, highlighting frameshift indels as a potential biomarker of checkpoint inhibitor response and supporting the targeting of clonal frameshift indels by both vaccine and cell therapy approaches.

Contributors

ST, KL, and CS designed the study and wrote the manuscript. ST, KL, and HX analysed the data. All authors interpreted the data.

Declaration of interests

ST reports grants from Ventana, outside the submitted work; and has a patent on indel burden and checkpoint inhibitor response pending, and a patent on targeting of frameshift neoantigens for personalised immunotherapy pending. KL has a patent on indel burden and checkpoint inhibitor response pending, and a patent on targeting of frameshift neoantigens for personalised immunotherapy pending. NM reports personal fees from Achilles Therapeutics, outside of the submitted work; and has two patents pending (PCT/EP2016/059401 and PCT/EP2016/071471). RR reports personal fees from Achilles Therapeutics, outside of the submitted work; and has two patents pending (GB1516047.6 and GB1601098.5). SAQ reports personal fees and other from Achilles Therapeutics, outside of the submitted work. JL reports personal fees from Eisai, GlaxoSmithKline, Kymab, Roche/Genentech, Secarna, Pierre Fabre, and EUSA Pharma; and grants and personal fees from Bristol-Myers Squibb, Merck Sharp & Dohme, Pfizer, and Novartis, outside of the submitted work. CS reports personal fees from Janssen, Boehringer Ingelheim, Ventana, Novartis, Roche, Sequenom, Natera, Achilles Therapeutics, and Sarah Cannon Research Institute, and personal fees and other from Apogen Biotechnologies, Epic Sciences, and GRAIL, outside of the submitted work; and has a patent on indel burden and checkpoint inhibitor response pending and a patent on targeting of frameshift neoantigens for personalised immunotherapy pending. All other authors declare no competing interests.

Acknowledgments

ST is funded by Cancer Research UK (grant reference number C50947/A). TC, ST, MG, and JL are funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at the Royal Marsden Hospital National Health Service Foundation Trust (ST grant reference number A109). KL is supported by a UK Medical Research Council Skills Development Fellowship Award (grant reference number MR/P014712/1). CS is funded by Cancer Research UK (TRACERx), the Rosetrees Trust, NovoNordisk Foundation (ID 16584), EU FP7 (projects PREDICT and RESPONSIFY, ID number 259303), the Prostate Cancer Foundation, the Breast Cancer Research Foundation, the European Research Council (THESEUS), and National Institute for Health Research University College London Hospitals Biomedical Research Centre. We would like to thank Zoltan Szallasi for his helpful review of the data and the manuscript. We thank Levi Garraway, Eli Van Allen, Alexandra Snyder, Matthew Hellman, Naiyer Rizvi, and Timothy Chan for kindly allowing access to their checkpoint inhibitor datasets. The results published here are in part based on data generated by The Cancer Genome Atlas Research Network.

References

1 Gubin MM, Artyomov MN, Mardis ER, Schreiber RD. Tumor neoantigens: building a framework for personalized cancer immunotherapy. *J Clin Invest* 2015; **125:** 3413–21.

2 Cancer Genome Atlas Research, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* 2013; **45:** 1113–20.

3 Yang H, Zhong Y, Peng C, Chen JQ, Tian D. Important role of indels in somatic mutations of human cancer genes. *BMC Med Genet* 2010; **11:** 128.

4 Tran E, Robbins PF, Rosenberg SA. 'Final common pathway' of human cancer immunotherapy: targeting random somatic mutations. *Nat Immunol* 2017; **18:** 255–62.

5 Gubin MM, Zhang X, Schuster H, et al. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* 2014; **515:** 577–81.

6 van Rooij N, van Buuren MM, Philips D, et al. Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *J Clin Oncol* 2013; **31:** e439–42.

7 Robbins PF, Lu YC, El-Gamil M, et al. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat Med* 2013; **19:** 747–52.

8 Prickett TD, Crystal JS, Cohen CJ, et al. Durable complete response from metastatic melanoma after transfer of autologous T cells recognizing 10 mutated tumor antigens. *Cancer Immunol Res* 2016; **4:** 669–78.

9 Tran E, Robbins PF, Lu YC, et al. T-cell transfer therapy targeting mutant KRAS in cancer. *N Engl J Med* 2016; **375:** 2255–62.

10 Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature* 2013; **500:** 415–21.

11 Linnemann C, van Buuren MM, Bies L, et al. High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4+ T cells in human melanoma. *Nat Med* 2015; **21:** 81–85.

12 Van Allen EM, Miao D, Schilling B, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* 2015; **350:** 207–11.

13 Snyder A, Makarov V, Merghoub T, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med* 2014; **371:** 2189–99.

14 Rizvi NA, Hellmann MD, Snyder A, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 2015; **348:** 124–28.

15 Rosenberg JE, Hoffman-Censits J, Powles T, et al. Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial. *Lancet* 2016; **387:** 1909–20.

16 Le DT, Uram JN, Wang H, et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N Engl J Med* 2015; **372:** 2509–20.

17 Senbabaoglu Y, Gejman RS, Winer AG, et al. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol* 2016; **17:** 231.

18 Gore ME, Griffin CL, Hancock B, et al. Interferon alfa-2a versus combination therapy with interferon alfa-2a, interleukin-2, and fluorouracil in patients with untreated metastatic renal cell carcinoma (MRC RE04/EORTC GU 30012): an open-label randomised trial. *Lancet* 2010; **375:** 641–48.

19 McDermott DF, Cheng SC, Signoretti S, et al. The high-dose aldesleukin "select" trial: a trial to prospectively validate predictive models of response to treatment in patients with metastatic renal cell carcinoma. *Clin Cancer Res* 2015; **21:** 561–68.

20 McDermott DF, Sosman JA, Sznol M, et al. Atezolizumab, an anti-programmed death-ligand 1 antibody, in metastatic renal cell carcinoma: long-term safety, clinical activity, and immune correlates from a phase IA study. *J Clin Oncol* 2016; **34:** 833–42.

21 Motzer RJ, Escudier B, McDermott DF, et al. Nivolumab versus everolimus in advanced renal-cell carcinoma. *N Engl J Med* 2015; **373:** 1803–13.

22 Linnebacher M, Gebert J, Rudy W, et al. Frameshift peptide-derived T-cell epitopes: a source of novel tumor-specific antigens. *Int J Cancer* 2001; **93:** 6–11.

23 Sato Y, Yoshizato T, Shiraishi Y, et al. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat Genet* 2013; **45:** 860–67.

24 Gerlinger M, Horswell S, Larkin J, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet* 2014; **46:** 225–33.

25 Hugo W, Zaretsky JM, Sun L, et al. Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell* 2017; **168:** 542.

26 Jamal-Hanjani M, Wilson GA, McGranahan N, et al. Tracking the evolution of non-small-cell lung cancer. *N Engl J Med* 2017; **376:** 2109–21.

27 McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20:** 1297–303.

28 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25:** 1754–60.

29 Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25:** 2078–79.

30 Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012; **22:** 568–76.

31 Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013; **31:** 213–19.

32 Cancer Genome Atlas Research N. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 2013; **499:** 43–49.

33 Lindeboom RG, Supek F, Lehner B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat Genet* 2016; **48:** 1112–18.

34 Roth A, Khattra J, Yap D, et al. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* 2014; **11:** 396–98.

35 Van Loo P, Nordgard SH, Lingjaerde OC, et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci USA* 2010; **107:** 16910–15.

36 Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* 2015; **160:** 48–61.

37 Shukla SA, Rooney MS, Rajasagi M, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat Biotechnol* 2015; **33:** 1152–58.

38 Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 2012; **28:** 1811–17.

39 Nielsen M, Lundegaard C, Blicher T, et al. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One* 2007; **2:** e796.

40 Jamal-Hanjani M, Wilson GA, McGranahan N, et al. Tracking the evolution of non-small-cell lung cancer. *N Engl J Med* 2017; **376:** 2109–21.

41 Giannakis M, Mu XJ, Shukla SA, et al. Genomic correlates of immune-cell infiltrates in colorectal carcinoma. *Cell Rep* 2016; **17:** 1206.

42 Fritsch EF, Rajasagi M, Ott PA, Brusic V, Hacohen N, Wu CJ. HLA-binding properties of tumor neoepitopes in humans. *Cancer Immunol Res* 2014; **2:** 522–29.

43 McGranahan N, Furness AJ, Rosenthal R, et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* 2016; **351:** 1463–69.

44 Nanda R, Chow LQ, Dees EC, et al. Pembrolizumab in patients with advanced triple-negative breast cancer: phase Ib KEYNOTE-012 study. *J Clin Oncol* 2016; **34:** 2460–67.

45 Nolan E, Savas P, Policheni AN, et al. Combined immune checkpoint blockade as a therapeutic strategy for BRCA1-mutated breast cancer. *Sci Transl Med* 2017; published online June 7. DOI:10.1126/scitranslmed.aal4922.

46 Streisinger G, Okada Y, Emrich J, et al. Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday. *Cold Spring Harb Symp Quant Biol* 1966; **31:** 77–84.

47 Furney SJ, Turajlic S, Stamp G, et al. The mutational burden of acral melanoma revealed by whole-genome sequencing and comparative analysis. *Pigment Cell Melanoma Res* 2014; **27:** 835–38.

48 Therkildsen C, Joost P, Lindberg LJ, Ladelund S, Smith-Hansen L, Nilbert M. Renal cell cancer linked to Lynch syndrome: increased incidence and loss of mismatch repair protein expression. *Int J Urol* 2016; **23:** 528–29.

49 Miuma S, Saldivar JC, Karras JR, et al. Fhit deficiency-induced global genome instability promotes mutation and clonal expansion. *PLoS One* 2013; **8:** e80730.

50 Cancer Genome Atlas Research N, Linehan WM, Spellman PT, et al. Comprehensive molecular characterization of papillary renal-cell carcinoma. *N Engl J Med* 2016; **374:** 135–45.

51 Davis CF, Ricketts CJ, Wang M, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* 2014; **26:** 319–30.

52 Scelo G, Riazalhosseini Y, Greger L, et al. Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat Commun* 2014; **5:** 5135.

53 Kuroda K, Hibi D, Ishii Y, et al. Ochratoxin A induces DNA double-strand breaks and large deletion mutations in the carcinogenic target site of gpt delta rats. *Mutagenesis* 2014; **29:** 27–36.

54 Margulis V, Lin J, Yang H, Wang W, Wood CG, Wu X. Genetic susceptibility to renal cell carcinoma: the role of DNA double-strand break repair pathway. *Cancer Epidemiol Biomarkers Prev* 2008; **17:** 2366–73.

55 Wang Y, Cortez D, Yazdi P, Neff N, Elledge SJ, Qin J. BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures. *Genes Dev* 2000; **14:** 927–39.

56 Maby P, Tougeron D, Hamieh M, et al. Correlation between density of CD8+ T-cell infiltrate in microsatellite unstable colorectal cancers and frameshift mutations: a rationale for personalized immunotherapy. *Cancer Res* 2015; **75:** 3446–55.

57 Chow S, Galvis V, Pillai M, et al. High-dose interleukin2—a 10-year single-site experience in the treatment of metastatic renal cell carcinoma: careful selection of patients gives an excellent outcome. *J Immunother Cancer* 2016; **4:** 67.

58 Geynisman DM. Anti-programmed cell death protein 1 (PD-1) antibody nivolumab leads to a dramatic and rapid response in papillary renal cell carcinoma with sarcomatoid and rhabdoid features. *Eur Urol* 2015; **68:** 912–14.

59 Ruiz-Banobre J, Anido U, Abdulkader I, Antunez-Lopez J, Lopez-Lopez R, Garcia-Gonzalez J. Long-term response to nivolumab and acute renal failure in a patient with metastatic papillary renal cell carcinoma and a PD-L1 tumor expression increased with sunitinib therapy: a case report. *Front Oncol* 2016; **6:** 250.

60 D'Angelo SP, Larkin J, Sosman JA, et al. Efficacy and safety of nivolumab alone or in combination with ipilimumab in patients with mucosal melanoma: a pooled analysis. *J Clin Oncol* 2017; **35:** 226–35.

61 Furney SJ, Turajlic S, Stamp G, et al. Genome sequencing of mucosal melanomas reveals that they are driven by distinct mechanisms from cutaneous melanoma. *J Pathol* 2013; **230:** 261–69.

62 Lappalainen T, Sammeth M, Friedlander MR, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 2013; **501:** 506–11.

63 Wang D, Zavadil J, Martin L, et al. Inhibition of nonsense-mediated RNA decay by the tumor microenvironment promotes tumorigenesis. *Mol Cell Biol* 2011; **31:** 3670–80.

64 Huang J, El-Gamil M, Dudley ME, Li YF, Rosenberg SA, Robbins PF. T cells associated with tumor regression recognize frameshifted products of the CDKN2A tumor suppressor gene locus and a mutated HLA class I gene product. *J Immunol* 2004; **172:** 6057–64.

65 Maby P, Galon J, Latouche JB. Frameshift mutations, neoantigens and tumor-specific CD8(+) T cells in microsatellite unstable colorectal cancers. *Oncoimmunology* 2016; **5:** e1115943.

66 Schwitalle Y, Kloor M, Eiermann S, et al. Immune response against frameshift-induced neopeptides in HNPCC patients and healthy HNPCC mutation carriers. *Gastroenterology* 2008; **134:** 988–97.

67 Williams DS, Bird MJ, Jorissen RN, et al. Nonsense mediated decay resistant mutations are a source of expressed mutant proteins in colon cancer cell lines with microsatellite instability. *PLoS One* 2010; **5:** e16012.

68 Le DT, Durham JN, Smith KN, et al. Mismatch-repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* 2017; published online June 8. DOI:10.1126/science.aan6733.

69 Lai J, Zhou Z, Tang XJ, Gao ZB, Zhou J, Chen SQ. A tumor-specific neo-antigen caused by a frameshift mutation in BAP1 is a potential personalized biomarker in malignant peritoneal mesothelioma. *Int J Mol Sci* 2016; published online May 14. DOI:10.3390/ijms17050739.

70 Wherry EJ, Kurachi M. Molecular and cellular insights into T cell exhaustion. *Nat Rev Immunol* 2015; **15:** 486–99.

71 Duan F, Duitama J, Al Seesi S, et al. Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J Exp Med* 2014; **211:** 2231–48.

72 Ruark E, Munz M, Clarke M, et al. OpEx—a validated, automated pipeline optimised for clinical exome sequence analysis. *Sci Rep* 2016; **6:** 31029.