Insights for precision healthcare from the 100,000 Genomes Cancer Programme

Nirupa Murugaesu (nirupa.murugaesu@genomicsengland.co.uk)
Genomics England

Alona Sosinsky
Genomics England

John Ambrose
Genomics England

William Cross
UCL Cancer Institute

Clare Tumbull
The Institute of Cancer Research

Shirley Henderson
Genomics England

J Jones
Barts Cancer Institute - CRUK Centre of Excellence

Angela Hamblin
Genomics England

Prabhu Arumugam
Genomics England

Georgia Chan
Genomics England

Daniel Chubb
Institute of Cancer Research

Boris Noyvert
University of Birmingham

Jonathan Mitchell
Genomics England

Susan Walker
Genomics England

Katy Bowman
Genomics England

Dorota Pasko
Genomics England

Marianna Pereira
Analysis

Keywords:

Posted Date: December 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2394107/v1
License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. The authors affiliated to Genomics England were employed or seconded to this company set up and owned by the United Kingdom Department of Health and Social Care.

Version of Record: A version of this preprint was published at Nature Medicine on January 11th, 2024. See the published version at https://doi.org/10.1038/s41591-023-02682-0.
Insights for precision healthcare from the 100,000 Genomes Cancer Programme

Alona Sosinsky1*, John Ambrose1*, Will Cross2*, Clare Turnbull3, Shirley Henderson1, Louise Jones1, Angela Hamblin1, Prabhu Arumugam1, Georgia Chan1, Dan Chubb3, Boris Noyvert, Jonathan Mitchell1, Susan Walker1, Katy Bowman1, Dorota Pasko1, Marianna Buongermino Pereira1, Nadezda Volkova1, Antonio Rueda-Martin1, Daniel Perez-Gil1, Javier Lopez1, John Pullinger1, Afshan Siddiq1, Tala Zainy1, Tasnima Choudhury1, Augusto Rendon1, Tom Fowler1, Sandra Hing4, Zandra Deans4, Genomics England Research Consortium, Sue Hill4, Mark Caufield5**, Nirupa Murugaesu1**

These authors contributed equally*, Corresponding authors#

Corresponding authors:
Mark Caufield; William Harvey Research Institute and the Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ, Telephone: +44(0)207 882 3402, email: m.j.caulfield@qmul.ac.uk

Nirupa Murugaesu; 1 Canada Square, Canary Wharf, London E14 5AA, Telephone: +44(0)207 882 5030, email: nirupa.murugaesu@genomicsengland.co.uk

Affiliation(s), and full address(es) of the authors
1. Genomics England, 1 Canada Square, London E14 5AA
2. UCL Cancer Institute, London WC1E 6DD
3. Institute of Cancer Research, 237 Fulham Road, London SW3 6JB
5. William Harvey Research Institute and the Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ
Abstract

The Cancer Programme of the 100,000 Genomes Project was a transformational UK government initiative that aimed to bring whole genome sequencing (WGS) to cancer patients and evaluate the opportunities for precision cancer care. Genomics England, in partnership with NHS England, generated whole genome analyses for 13,880 solid tumours across 33 different cancer types, and genomic data were linked with real-world health data within a secure national research environment. Here, we report the overall findings of the programme, focusing on clinical actionability and potential wider clinical significance. We found variation between cancer types in the incidence of somatic mutations of different types in genes currently recommended for standard-of-care testing. For example, 94% of glioblastoma multiforme cases had small variants and 54% had copy number aberrations (CNAs) in at least one gene recommended for clinical testing, whereas sarcoma was found to have the highest proportion of actionable structural variants (13%). We confirmed the importance of utilising pan-genomic markers, such as mutational signatures, with 51% of high grade serous ovarian cancer cases showing homologous recombination deficiency, 13% of which were associated with pathogenic germline variants, indicating the value of combined somatic and germline analyses. We also observed a significant co-occurrence of somatic small variants and CNAs in several known oncogenes including \textit{EGFR, GNAS, BRAF, and KRAS}. Our findings demonstrate the value of combining genomic testing with real world clinical and treatment data to inform clinical recommendations for genomic testing in cancer, to enable survival analysis and improve understanding of the long-term effects of clinical cancer genomics on patient outcomes.
Over the last decade, the incidence rate for all cancers combined has increased by around 4% in the UK\(^1\). Therefore, there is an increased need for molecular cancer testing allied with established and emerging targeted therapies, alongside testing of cancer predisposition genes and pharmacogenomic markers\(^2\). The 100,000 Genomes Project was a transformational UK government initiative, conducted within the National Health Service (NHS) in England. The primary goal was to undertake whole genome sequencing (WGS) for patients with cancer, rare disease and infectious diseases and standardise high throughput, automated bioinformatic analysis and interpretation\(^3\). Participants gave consent for their genomic data to be linked to anonymised longitudinal health records and shared with researchers in a secure workspace to drive forward our knowledge of the causes and treatment of different cancers\(^4\).

The aim of the Cancer Programme of the 100,000 Genomes Project was to develop the infrastructure for routine high-throughput sequencing and to evaluate the role of WGS at scale for NHS cancer patients (Figure 1A). In parallel, the data generated were used to establish a national research platform of molecular data (National Genomic Research Library) with secure linkage to longitudinal real-world data. The national clinical datasets incorporated include the National Cancer Registration and Analysis Service Dataset (NCRAS) consisting of cancer registration data and the Systemic Anti-Cancer Therapy Dataset (SACT), as well as subsequent cancer episodes including, Hospital Episode Statistics (HES) and mortality data from the Office of National Statistics (ONS)\(^5\) (Figure 1B).

The longer term objective was to accelerate the delivery of molecular testing in NHS clinical cancer care\(^6\). Building on the work of the 100,000 Genomes Project, on 1 January 2019, the NHS Genomic Medicine Service (GMS) was launched to deliver genomic testing for rare diseases and cancer across England, using a standardised National Genomic Testing Directory, including provision for single-gene tests, targeted large gene panels and WGS. The National Genomic Testing Directory aims to provide consistency of test methodologies across clinical indications in healthcare via a consolidated network of seven NHS England Regional Genomic Laboratory Hubs\(^7\). It specifies the genomic tests that are commissioned and thereby funded by the NHS in England as part of gold-standard molecular profiling to guide cancer care across different cancer clinical indications. In addition, the NHS GMS provides further opportunities for patients to participate in research\(^8\).
Whilst large-scale sequencing studies such as the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) have extensively catalogued the spectra of somatic mutations present in different solid tumour types\(^9\)-\(^11\), here, we present the results from WGS data, focused on clinically actionable genes and pan-genomic markers, linked to real world longitudinal, life-course clinical data and highlight the learnings from the Cancer Programme and the implications for this in current clinical care.

**Results**

**Cohort demographics.** We sequenced 16,358 tumour-normal sample pairs from 15,241 individuals that were recruited to the Cancer Programme of the 100,000 Genomes Project between 2015 and 2019 (Figure 1A-B, recruitment timeline shown in Supplementary Figure 1A). Here, we describe the integrative whole genome analysis for 33 tumour types of 13,880 tumour samples (13,311 (95.9%) derived from fresh frozen tumour tissue and 569 (4.1%) from formalin-fixed paraffin-embedded tumour tissue) and their matching normal (germline) samples (13,493 (99.1%) blood-derived, 100 (0.7%) from normal tissue and 23 (0.2%) from saliva samples (Figure 1C). Genomes for haematological tumours (841), paediatric cancers (333, less than 18 years of age), carcinomas of unknown primary (98) and tumours that were not linked to external datasets (1,206) were excluded from this analysis. Tumour types with more than 1,000 sequenced tumour genomes include breast invasive carcinoma (n=2,861), colon adenocarcinoma (n=1,887), kidney renal clear cell carcinoma (n=1,073) and sarcoma (n=1,026). The number of patients recruited nationally by 13 participating Genomic Medicine Centres (comprising over 80 Hospital Trusts) across England are shown in Figure 1D. The distribution of age and gender across the different tumour types is shown in Supplementary Figures 1B and 1C. Early onset with median age <50 years was demonstrated for low grade glioma and testicular germ cell tumours in agreement with incidence statistics\(^12\).

Staging information was available in the NCRAS dataset for 12,040 tumours (12,040/13,880, 86.7%) and the breakdown of the different stages of the tumour types sequenced is demonstrated in Figure 1E. 11.9% (1,645/13,880) of patients had stage 4 cancer (advanced metastatic disease) with specimens obtained from metastatic sites including the liver, lymph nodes, lung and brain (Figure 1E). We found a higher prevalence of patients with advanced disease (stage 3 and 4) amongst high grade serous ovarian carcinoma and skin cutaneous melanoma, whilst in other tumour types, such as invasive breast cancer, a higher prevalence of patients with early stage (stage 1 and 2) disease were recruited as a consequence of biases in tissue ascertainment. The majority of tumour samples were obtained from surgical resections (94.5%, 13,120/13,880, Figure 1E) consisting of 93.6% treatment-naive, surgical resections
(12,275/13,120) and 6.4% resections following neoadjuvant treatment (845/13,120). This is compared with only 5.5% (760/13,880) from metastatic or diagnostic, biopsy samples, and of those, 10.9% (83/760) were post-treatment (Figure 1E). The tumour content for sequenced samples is shown in Figure 1G. This highlights potential difficulties in collecting samples with sufficient tumour content (>30%) in specific tumours such as lung adenocarcinoma and lung squamous cell carcinoma, head and neck squamous cell carcinoma and pancreatic adenocarcinoma, consistent with previous publications.

**Clinical actionability through WGS.** A single test such as WGS can facilitate the concurrent detection of somatic small variants (including single nucleotide variants (SNVs) and insertions and deletions (indels)), copy number aberrations (CNAs) and structural variants (SVs), including gene fusions. In addition, when paired sequencing of both tumour and normal samples is performed, germline findings such as variants in cancer susceptibility genes and pharmacogenomic findings (variants affecting the metabolism of therapeutic agents used to treat cancers) can also lead to a greater yield of clinically relevant findings. Through the Cancer Programme, whole genome analysis (WGA) results generated in an automated bioinformatics pipeline were returned to Genomic Medicine Centre Laboratories using a standardised format, and potentially actionable findings were reviewed initially by clinical scientists and subsequently at multi-disciplinary Molecular Tumour Boards, referred to as Genomic Tumour Advisory Boards (GTABs). Examples of WGA results are shown in Supplementary Figure 1E and full details of analysis and interpretation are described in the Methods section, showing the utility of WGS to capture various genomic alterations of clinical significance with a single test.

We assessed the findings from the aggregated data from the 100,000 Genomes Project for solid cancers in the context of the current National Genomic Test Directory for Cancer (NGTDC). The results across the cohort of 13,380 whole genomes revealed various types of mutations relating to targets specified in the NGTDC, including small variants, CNAs and fusions, along with germline variants associated with inherited cancer risk and pharmacogenomic findings (see Online Methods for details). We found that the percentage of cases with one or more somatic mutation present in genes indicated in the NGTDC for the applicable cancer type was high, though variable (Figure 2). For example, more than 50% of tumours harboured one or more mutation found in genes indicated for testing in the NGTDC in glioblastoma multiforme, low grade glioma, skin cutaneous melanoma, head and neck squamous cell carcinoma, colon and rectal adenocarcinoma, lung adenocarcinoma and ovarian high grade serous carcinoma (Figure 2). Clinically relevant mutations were found in 20-49% of breast invasive carcinoma, sarcoma, mesothelioma,
uterine endometrial, and lung squamous cell carcinoma cases, while in other cancer types such as pancreatic, prostate, esophageal and stomach adenocarcinoma less than 20% of cases possessed mutations in genes present in the NGTDC (Figure 2). We note that clinical actionability of these mutations will be dependent on the individual case, based on clinical circumstances such as the stage of the tumour and associated co-morbidities of the participant recruited to the study. This, therefore, highlights the need for clinical interpretation and discussion where clinically appropriate within a GTAB.

We also examined mutations listed in the NGTDC in other cancer types for which testing of that gene/mutation is not currently indicated (Figure 2 and in Supplementary Figure 2A, 2B and 2C). These variants are denoted in blue within Figure 2 and could indicate potentially actionable findings that may enable recruitment into clinical trials or prompt further review within a GTAB. For example, SNVs were identified in PIK3CA and KRAS across different cancer types and similarly pan-genomic markers, such as the HRD signature, for which clinical trials may be available. We envisage that the clinical indications within the NGTDC will continue to expand as new evidence from biomarker-driven trials becomes available, and genes and biomarkers will be extended to new cancer types as well as novel biomarkers being included.

**Landscape of somatic small variants.** The most frequently mutated gene was TP53 (5,411/13,880, 39.0% of patients harbouring predicted, protein-altering SNVs or indels; Figure 2A and Online Methods for details on reported variants), with TP53 mutations found in more than 70% of cases in uterine corpus endometrial serous carcinoma, ovarian high grade serous carcinoma, lung squamous cell carcinoma, rectum adenocarcinoma, esophageal adenocarcinoma and esophageal squamous cell carcinoma, but lower in other cancer types. Of individuals with at least one TP53 mutation, 36% (1959/5411) carried one or more variant predicted to be protein-truncating or splice-altering and 65.5% (3544/5,411) carried one or more missense variant (207 individuals carried both variant types), with the five most common protein changes being R175H, 5.3%; R273C, 3.2%; R248Q, 3.2%; R273H, 3.2% and R282W, 2.7%. PIK3CA was the second most frequently altered gene, with mutations found in 19.8% of patients (2,750/13,880), occurring most frequently in uterine corpus endometrial carcinoma (53.5%), ovarian endometrioid adenocarcinoma (49.0%), breast invasive carcinoma (42.2%), uterine corpus endometrial serous carcinoma (38.1%) and colon adenocarcinoma (26.5%). The most commonly mutated codons were PIK3CA H1047, PIK3CA E545 and over 90% comprised five well-characterised hotspots in PIK3CA. The next most commonly mutated genes were APC (18.4%) and KRAS (12.8%), although these are enriched in a few cancer types including
colon, rectal and lung adenocarcinoma. Although some genes were mutated at similar rates across many
tumour types (for example, \textit{TP53} and \textit{PIK3CA}), others such as \textit{APC}, \textit{VHL} and \textit{IDH1} were highly enriched for
mutations in only one or two tumour types. We note that these results are comparable to other pan-
cancer WGS studies\textsuperscript{16}, with differences in frequency consistent with the increased proportion in our series
of certain cancer types, such as colon and rectal adenocarcinoma and sarcoma in our cohort (Figure 1C).

\textbf{Fusions and copy-number aberrations.} The detection of fusions and CNAs has expanded over the past
few years with associated targeted therapies across cancer types\textsuperscript{17}, and is now an essential part of
standard-of-care testing. For instance, there are now a number of different agents available for targeted
treatment of tumours that harbour \textit{NTRK} (across all cancer types) and other kinase fusions (e.g. \textit{ALK}, \textit{ROS},
\textit{RET} for lung cancers). These targets are included in the NGTDC, though only a small percentage of patients
test positive for each individual fusion. Fusions that are essential for confirming the classification of
sarcoma tumours (e.g. \textit{HEY1-NCOA2} in mesenchymal chondrosarcoma) are also included into the NGTDC
and make significant contributions to refining diagnoses and identifying actionable findings for this
tumour type. In keeping with this, sarcomas had the highest percentage of tumours (13\%) with clinically
relevant structural variants findings\textsuperscript{18}. We found that CNAs, defined as copy gains where copy number
was at least twice higher than the overall ploidy for oncogenes or losses defined as complete copy loss or
loss of heterozygosity combined with a non-synonymous small variant in the same gene for tumour
suppressor genes (See Online Methods), were very common across all cancer types and affected cancer-
specific genes noted in the NGTDC. Notably, for certain cancer types including glioblastoma multiforme,
low grade glioma, head and neck squamous cell carcinoma, mesothelioma and sarcoma (Figure 2 and
Supplementary Figure 2B), we found CNAs in nearly all NGTDC genes, most commonly in \textit{TP53}, \textit{CDKN2A},
\textit{MYC}, \textit{CDKN2B}, \textit{PTEN} across all cancer types (Figure 2).

\textbf{Germline findings.} The median age at tumour diagnosis is shown in Supplementary Figure 1C, and as
expected, there is a younger median age of presentation for a tumour diagnosis in those patients with
predisposing germline findings (predicted truncating small variants or missense mutations with
pathogenic or likely pathogenic classification in Clinvar with a rating of two stars or higher; for details see
Online Methods) (Supplementary Figure 1D). Notably, patients with germline variants in MMR genes show
significantly earlier age of onset of colon adenocarcinomas, while patients with germline variants in HRD
genes show significantly earlier onset in ovarian high grade serous carcinomas and invasive breast cancer.
This was also observed in kidney renal clear cell carcinoma with germline variants predominantly in the
VHL gene. We evaluated the presence of variants in DPYD associated with toxicity to fluoropyrimidines and observed between 5-10% of cases carried these variants. In the NGTD this would lead to recommendations to omit or adjust the dose of fluoropyrimidines that are routinely given in the treatment of invasive breast, colon, rectal, pancreatic adenocarcinomas and head and neck squamous cell carcinomas.

Pan-genomic markers and mutational signatures. Tumour mutation burden (TMB), has been cited as a potential biomarker19, and we observed significant variation, both across and within cancer types. In line with previous reports20 we found that skin cutaneous melanoma and lung cancer had the highest average TMB (Figure 3A). There was considerable variability in both colorectal and uterine endometrial cancers, representative of the presence or absence of microsatellite instability (MSI) or hypermutation caused by POLE mutations (see alignment with corresponding mutational signatures).

When examining mutational signatures with well described aetiologies, we observed expected frequencies of certain mutational signatures associated with certain cancer types21. Over 50% of ovarian high grade serous carcinoma showed the presence of the homologous recombination deficiency (HRD) signature (Figure 3A). As expected, smoking and APOBEC signatures were associated with lung cancer, and UV signature with skin cutaneous melanoma and we found an enrichment of the DNA mis-match repair deficiency (MMR) signature in MSI colorectal and uterine endometrial cancers (Figure 3A).

Clinical utility of WGS. Overall, these findings demonstrate the ability of WGS data to assess SNVs, CNAs and gene fusions along with pan genomic markers (such as mutational signatures and TMB) and germline findings all through the utilisation of a single test, to fully characterise the clinical genomic landscape of a tumour (Figure 2). In Supplementary Figure 1E and 1F, we provide several examples of whole genome analysis (WGA) results as they were shared with Genomic Medicine Centre Laboratories, demonstrating the identification of clinically relevant genomic alterations. In an example of WGS results for ovarian high grade serous carcinoma, we identified a somatic TP53 SNV and TP53 copy number loss, consistent with the diagnosis, as well as a germline BRCA1 mutation and somatic BRCA1 copy number loss driving HR deficiency that is supported by mutational signature analysis. Similarly, in another example report of colorectal cancer, we identified the presence of MMR deficiency signature and high TMB along with a PMS2 pathogenic germline variant, somatic PMS2 and MSH6 mutations and a pharmacogenomic (germline) variant in the DPYD gene (associated with toxicity to fluoropyrimidines).
Pan-genomic markers and outcomes from real world data. Through the linkage of the WGS data with longitudinal life course clinical data (Figure 1B) it was possible to assess treatment outcomes for patients stratified by pan-genomic markers. We found, as expected, that the majority of cases treated with platinum chemotherapy in the presence of HR deficiency (n=254) were invasive breast cancers (n=59, 22%) and ovarian high grade serous carcinomas (n=161, 60%). In cases treated with immunotherapy in the presence of MMR deficiency (n=12), colon adenocarcinoma (n=7, 58%) was the most frequently treated cancer type. As shown in Figure 3B, HR deficiency signatures predict outcome for those patients treated with platinum therapies (p < 0.0001). This was not significant in MMR deficient cases treated with immunotherapies although the number of cases was small (n=12).

We then evaluated TMB as a predictive marker in patients who received immunotherapy treatment\(^{22}\), which was predominantly in skin cutaneous melanoma and lung adenocarcinoma. Indeed, the highest median TMB level across all cancer types was observed in skin cutaneous melanoma (Figure 3A). A significant difference in survival (p=0.015, HR=2.38) was observed for those patients with TMB in the lowest quartile (median of 3.8 non-synonymous small variants/Mb) compared with the highest quartile (median of 20.98 non-synonymous small variants/Mb) in those diagnosed with skin cutaneous melanoma and treated with immunotherapies (Figure 3C). Interestingly, a significant difference was not observed in lung adenocarcinoma (p=0.72), where the lowest and highest quartile median TMB values were 2.2 and 10.5 non-synonymous small variants/Mb, respectively. This may indicate that the level of TMB is relevant in predicting response and supports the need for further refining of TMB as a predictive marker for immunotherapy response as highlighted in previous studies\(^{23,24}\).

Co-occurrence of copy-number alterations and cancer genes and associations with outcomes from real world data. The co-occurrence of somatic small variants (SNVs and indels) and CNAs is well documented\(^{25}\). With WGS, we were able to explore the co-occurrence of CNAs and somatic small variants impacting the same gene across cancer genes in the NGTDC. We divided cases with appropriate data (n=12,072) into those with and without small variants for each gene and then compared the frequency of cases with copy number alterations for each gene across the two groups (Figure 4A). Following multiple testing corrections, we found that 17 genes (45%) displayed significant differences in specific copy status. With these data, we confirmed several previous findings, namely, that \(\text{EGFR}\)\(^{26}\) and \(\text{KIT}\)\(^{27,28}\), in specific cancer types, tended to be amplified when a putative activating SNV was present. The role of copy gains on certain oncogenes has long been debated and our analysis found that there was a significant co-
occurrence of gains in the presence of **BRAF**, **KRAS**, and **FGFR2** small variants. We also found that 10 tumour suppressor or dual role genes had significantly higher frequencies of copy loss, or loss of heterozygosity when somatic small variants were present, including established examples such as **TP53**, **RB1**, **CDKN2A**, **BRCA1**, **NF1**, **APC**, **ATM**, **SDHA**, and **ATRX**. Together this supports the premise that WGS allows the copy states to be reviewed in actionable driver genes, again highlighting the value of being able to observe different types of mutations in parallel.

**Survival analysis using real world data.** We next examined overall survival across multiple tumour types stratified by presence or absence of mutations in cancer genes indicated for testing in the NGTDC. Together with protein altering small variants, we also assessed for copy number losses in tumour suppressor genes due to previous reports of haploinsufficiency\(^{29,30}\), within **TP53**\(^{31}\), **PTEN**\(^{32}\), **RB1**\(^{33}\), **ATRX**\(^{34}\), **NF1**\(^{35}\), **ATM**\(^{36}\) and **CDKN2A**\(^{37}\). We performed Kaplan-Meier and Cox-proportional hazard analyses on our pan-cancer cohort that included 33 cancer types. After correcting for stage and multiple testing, we found that alterations in 16 genes affected overall survival (Figure 4B). The gene that affected the outcome most severely is **CDKN2A** (p=7e-31, HR=2.3) which corresponds to its association with high grade disease and poor prognosis in some cancer subtypes, such as glioma\(^{38}\) and soft tissue sarcoma\(^{39}\). Our results agree with previously reported prognostic associations for specific tumour types, for example, poor prognosis for **KRAS** mutants in colorectal cancer\(^{40}\) and non-small cell lung cancer\(^{41}\) or **TP53** mutants in non-small cell lung cancer\(^{42}\). Mutations in **APC**, **ATM**, **BRCA1**, **BRCA2**, **CDK12**, **PIK3CA**, **MSH2**, **MSH6** and **POLD1** were associated with favourable outcomes, in keeping with reports in the literature for a number of these genes. For example, mismatch repair deficiency (caused by loss of function mutations in **MSH2** and **MSH6** genes) was associated with favourable prognosis in colorectal cancer due to immune escape\(^{43}\). Similarly, homologous recombination deficiency (that can be caused by loss of function mutations in **BRCA1** and **BRCA2** genes) has been associated with favourable prognosis\(^{44}\) in patients with triple negative breast cancer.

**Discussion**

Through the 100,000 Genomes Project, the infrastructure and resources to enable genomic data to be linked to clinical and longitudinal life course data within a healthcare setting, was established. Our findings from the Cancer Programme, support the selection of targets that have been indicated in the National Genomic Test Directory and demonstrate the value of identifying different types of mutations in the molecular characterisation of different cancers. Evaluation of WGS data has provided support for the
commissioning of WGS testing for cancers such as sarcoma, glioblastoma, high grade serous ovarian and triple negative breast cancers, in order to detect a range of molecular targets, including pan genomic markers, with a single test to inform clinical care. As a result of the 100,000 Genomes Project, a new NHS Genomic Medicine Service was created in England, that has standardised the molecular characterisation of tumours. In addition, standardised whole genome analysis reports for cancer are generated in an ISO-accredited bioinformatics pipeline and national guidance for the interpretation and reporting of somatic cancer variants, technical validation and return of predisposing germline findings have been established through NHS England-led national working groups. Findings are reviewed at multi-disciplinary Molecular Tumour Boards, or GTABs, which have been set up in federated networks across regions within the NHS in England. As more patients have genomic testing, it will be increasingly important to capture genomic data linked with real world longitudinal clinical and treatment data to progress our understanding of the long-term effects of clinical cancer genomics on patient outcomes.

As the capacity to sequence tumours expands, a core aim of the NHS Genomic Medicine Service is to extend the clinical benefit of prospective molecular characterisation to more patients with cancer. We have demonstrated that the longitudinal capture of linked real world clinical and treatment data allows evaluation of patient outcomes, and our findings recapitulate molecular markers of adverse outcomes seen in randomised clinical trials. The accumulation of genomic data alongside clinical data may further refine the selection of biomarkers, incorporating the co-occurrence of variants in the same gene or the coexistence of mutations in different genes, to enhance the prognostic and predictive value and may detect longer term signals of benefit or harm and aid clinical and regulatory decision making. The implication for therapeutic responses in the co-occurrence of CNAs alongside somatic small variants is not well known, and this level of genomic information may not readily be available from large cancer panel data. We present a broad survival analysis at the gene level and as the dataset expands it will be possible to look at specific variants within these cancer genes to fully examine the functional consequence of specific variants as prognostic markers, as previously observed with KRAS variants.

In parallel, molecular profiling initiatives will continue to grow as an important component of cancer care, as no single test is likely to be sufficient to address the many clinical needs. Whilst genomics may provide an important starting point for molecular stratification of cancer, it is likely that other modalities such as cell-free DNA, RNA sequencing, gene-expression profiling, proteomics, long-read sequencing and single-cell sequencing will continue to mature toward clinical utilisation. It is highly likely that we will need to
integrate multimodal molecular data, along with clinical data including digital pathology and radiology to maximise the benefit of precision cancer care for patients\textsuperscript{48,49}.

Challenges remain in implementing broad-based clinical WGS in the NHS in England. Providing a cutting-edge UK genomics service requires not only the sequencing and informatics infrastructure, but an expansion of operational requirements to aid local transformation, engagement and development of skills of the workforce and further expansion of the professional networks that have evolved through delivery of the 100,000 Genomes Project\textsuperscript{9}.

WGS was initially available in the NHS GMS for a small number of cancer types of clinical need, for example, sarcomas and paediatric tumours. As the cost of WGS continues to fall, and as the number and complexity of ‘actionable’ markers increases, there is expectation of successive tumour types moving to WGS as the standard test, in preference to gene panels. Our findings have facilitated adoption of WGS in the NGTD for metastatic solid cancers where standard-of-care testing and treatment options have been exhausted, where there is diagnostic uncertainty and for rare tumour types where there is limited molecular testing available. Our findings demonstrate the clinical utility of WGS in cancer and as it becomes more economical, it will likely become more widely used\textsuperscript{8}.

Through the NHS GMS, WGS results for cancer are discussed at a multi-disciplinary GTAB to ensure further appraisal of somatic and germline variants, determine clinical actionability and provide clinical recommendations. The role of GTABs includes ensuring that the treating multi-disciplinary team and clinicians are notified appropriately of actionable results, and eligibility for approved therapies and/or trials options are fully explored\textsuperscript{50}. A well-designed GTAB contributes to the clinical interpretation of cancer genomic testing, guides clinicians in their decision-making through recommendations, facilitates clinical trial enrolment and, may improve outcomes\textsuperscript{51,52}. This is in line with adaptive, basket trials such as DETERMINE\textsuperscript{53} which has been established to evaluate licensed treatments in non-licensed indications similar to the Drug Rediscovery Protocol (DRUP Trial)\textsuperscript{54,55}. The aim is to enable more equitable and comprehensive molecular testing within the NHS and to optimise cancer care by identifying all clinically relevant mutations for a specific cancer (as shown in Figure 2), but also to ensure patients are fully considered for clinical research and trials as a result of this genomic testing.
The National Genomic Research Library (NGRL), a platform built by Genomics England and NHS England, allows approved researchers to access samples, genomic data, and other associated health data via a secure Research Environment. Our findings highlight how these data can provide further information in terms of prognostic outcomes based on the absence or presence of a specific mutation. The accumulation of further genomic data within the NGRL, linked to clinical and outcome data, will enable more refined analyses and research to be undertaken using real world data\textsuperscript{56}. With more comprehensive profiling of tumours, it will be possible to further refine prognostic and predictive molecular markers, comprising not only combinations of different genomic alterations, but beyond genomics, and expand the reach of precision oncology to improve cancer outcomes.
Acknowledgements

We express our thanks to the participants, who made this work possible. Genomics England and the 100,000 Genomes Project was funded by the National Institute for Health Research, the Wellcome Trust, the Medical Research Council, Cancer Research UK, the Department of Health and Social Care and NHS England. We are grateful for the support from Professor Dame Sue Hill and the team in NHS England for establishing and funding the 13 Genomic Medicine Centres. We thank all of the NHS Genomic Medicine Centres located across England in the East of England, Imperial College and West London, Greater Manchester, North Thames, Northwest Coast, North East and Cumbria, Oxford, South Thames, South West, West of England, Wessex, West Midlands, Yorks and Humber. They enabled the NHS contribution including the clinical return of results within the NHS in a standardised and validated format. Mark Caulfield was an NIHR Senior Investigator between 2013-2021. This work is part of the portfolio of translational research at the NIHR Biomedical Research Centres at Barts, Cambridge University Hospitals NHS Foundation Trust, Great Ormond Street Foundation NHS Trust, Manchester University NHS Foundation Trust, Newcastle Hospitals NHS Foundation Trust, Oxford University Hospitals NHS Foundation Trust, Guys and St Thomas’s NHS Foundation Trust and University College London NHS Foundation Trust. We thank the NHS Genomic Laboratory Hubs and those who worked to create the National Genomic Test Directory, undertake its annual review and the NHS Genomic Medicine Service. This work was made possible through the generosity of NHS patients and uses clinical data from the NHS and NHS Digital. We thank all the staff at Genomics England Ltd. We gratefully acknowledge the Illumina Laboratory Services team at Hinxton for their advice and for undertaking the whole genome sequencing.
References


56. How your data is used. *Genomics England* https://www.genomicsengland.co.uk/patients-participants/data.

57. 100,000 Genomes Project Bioresource - main phase. *Health Research Authority* https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/100000-genomes-project-bioresource-main-phase/. 

Online Methods

Ethics
Approval from the national research ethics committee was obtained (IRAS ID 166046)\textsuperscript{57}. Participants were selected on the basis of having been identified by health care professionals and researchers within the NHS as having a cancer diagnosis. The participants were recruited across 13 NHS Genomic Medicine Centres and written informed consent was obtained from the participants.

Sample collection
Sample collection and DNA extraction requirements are described in Sample Handling Guidance (v4.0, available \url{https://files.genomicsengland.co.uk/forms/Sample-Handling-Guidance-v4.0.pdf}). 10 μg of germline DNA and at least 1.3μg of tumour DNA were required for Illumina TrueSeq PCR-free library preparation to be performed. PCR-based library preparation was employed when insufficient DNA could be obtained for PCR-free sequencing, with a minimum requirement of 500 ng. Optimised formalin fixed tumour tissue was allowed for WGS under exceptional circumstances, where the tumour size limited availability of fresh tissue, or if no tumour was present in the fresh frozen sample.

Analytical bioinformatics pipeline
For full details of the bioinformatics pipeline see the Cancer Genome Analysis Technical Information at \url{https://www.genomicsengland.co.uk/initiatives/100000-genomes-project/documentation}

Quality of sequencing data: All samples were sequenced on the HiSeq platform to an average coverage of 100x for tumour and 30x for normal. The following checks were implemented to ensure sample quality:

- normal samples had > 85 Gb and tumour samples had > 210 Gb of high quality sequencing data (base quality > 30, duplicated reads removed); normal samples had > 95% of the autosomal genome covered at ≥ 15x after removing reads with mapping quality < 10
- normal samples had cross-patient contamination < 3% as assessed by VerifyBamID
- tumour samples had cross-patient contamination < 2.5% and normal-tumour sample pair originating from the same patient as assessed by ConPair
- quality of sequencing data was monitored using PCA analysis based on the following metrics: percentage of reads mapped to the reference genome, proportion of chimeric DNA fragments,
median fragment size, unevenness of local genome coverage, percentage of reads missing from AT-rich or GC-rich genomic regions (AT and GC drop).

**Mapping and variant calling:** the Illumina North Star pipeline (version 2.6.53.23) was used for primary WGS analysis. Read alignment against the human reference genome GRCh38+Decoy+EBV was performed with ISAAC (version iSAAC-03.16.02.19). We acknowledge deficiencies in the ISAAC alignment software for precise variant allele frequency estimate [https://www.biorxiv.org/content/10.1101/836171v1.full](https://www.biorxiv.org/content/10.1101/836171v1.full) and for tumour evolution analysis and note that all 100K genomes were recently realigned with the Illumina Dragen platform (data available in the National Genomics Research Library). Small variant calling together with tumour-normal subtraction was performed using Strelka (version 2.4.7). In addition to default Strelka filters we applied the following additional filters in order to reduce the false positive rate in the final set of somatic variants: i) Variants with a population germline allele frequency above 1% in the Genomics England or gnomAD datasets; ii) Recurrent somatic variants with frequency above 5% in the Genomics England dataset; iii) Variants overlapping simple repeats as defined by Tandem Repeats Finder; iv) Small indels in regions with high levels of sequencing noise where at least 10% of the basecalls in a window extending 50 bases to either side of the indel call were filtered out by Strelka due to poor quality; v) SNVs resulting from systematic mapping and calling artefacts with a Fisher’s exact test phred score < 50. We tested whether the ratio of tumour allele depths at each somatic SNV site were significantly different to the ratio of allele depths at this site in a panel of normals (PoN). The PoN was composed of a cohort of 7000 non-tumour genomes from the Genomics England dataset, and at each genomic site only individuals not carrying the relevant alternate allele were included in the count of allele depths.

Copy number aberrations were identified with Canvas 1.3.1. Manta (version 0.28.0) was used to call Structural Variants (SVs) and long indels (>50bp), combining paired and split-read evidence for SV discovery and scoring.

**Annotation and reporting actionability:** SNVs and small indels were left-aligned, trimmed, and multi-allelic variants decomposed, prior to annotation with Cellbase [22693220], using Ensembl (version 90/GRCh38), COSMIC (version v86/GRCh38) and ClinVar (October 2018 release) databases. Annotation of consequence types was carried out by a high-performance variant annotator within Cellbase, and only variants annotated with a curated set of consequence types (stop gained/lost, start lost, frameshift
variant, inframe insertion/deletion, missense variant, splice acceptor/donor variant, splice region variant) in canonical transcripts were reported.

Interpretation of Copy Number Aberrations (CNAs) took into account gene mode of action as defined in COSMIC Cancer Gene Census (i.e. oncogene or tumour suppressor gene). Where a gene had an ambiguous or unknown role in cancer, it was included in both oncogene and tumour suppressor categories. Gains in oncogenes were reported if copy number was at least twice higher than the overall ploidy as defined by Canvas. The following scenarios were reported as losses in tumour suppressor genes: i) homozygous deletions called by Canvas (copy number, CN=0), ii) loss of heterozygosity called by Canvas (CN=1) in combination with a non-synonymous somatic small variant, and iii) Manta SVs with the potential to disrupt the gene coding region. Only samples with tumour purity >30% were included in the CNA actionability analysis.

Manta calls (BND, DEL, DUP or INV) were further assessed for the potential to generate productive fusions using an in-house approach based on transcript orientation and consistency of reading frame across the SV breakpoint. SVs that were identified as out-of-frame or untranscribed were discarded. Potential in-frame fusions and ambiguous events with a breakpoint in the coding exon or in the 5'UTR of downstream partners were reported.

Germline variants listed in ClinVar as pathogenic or likely pathogenic with a rating of at least two stars and predicted protein truncating variants in the genes for which the mechanism of pathogenicity is loss of function were reported for a subset of cancer predisposition genes.

**Signatures and tumour mutation burden:** For each tumour sample, frequencies across all SNV trinucleotide contexts were calculated and the contribution of each of the COSMIC (v2) signatures to the overall mutation burden observed in the tumour was derived using decomposition by non-negative least squares. Tumour mutation burden (TMB) was calculated as the total number of non-synonymous somatic small variants per megabase of coding sequence.

**Description of clinical data resources**
A minimal set of patient and sample data was collected from Genomic Medicine Centres (GMCs) at the time of DNA sample submission (e.g. tumour type, year of birth, gender, tissue source). Secondary clinical
information was gathered from NHS Digital (NHSD) and Public Health England (PHE)/National Cancer Registration and Analysis Service (NCRAS). From NHSD, Hospital Episode Statistics (HES) data was used to obtain details of all commissioned activity during admissions, and mortality information was obtained from Office of National Statistics (ONS) registry data for cancer registrations and deaths inside and outside of hospitals. From PHE/NCRAS, the av_tumour table was used to obtain tumour date of diagnosis, together with histology and morphology codes and the Systemic Anti-Cancer Therapy (SACT) table provided information on date and types of treatment. All datasets were accessed via the National Genomics Research Library using LabKey.

**Linking genomic data with secondary data sources**

Haematological tumours, paediatric tumours and carcinomas of unknown primary were considered to be outside the scope of the study and were removed prior to tumour selection. Secondary data from Public Health England/National Cancer Registration and Analysis Service (NCRAS) tumour catalogue (av_tumour), and NHS Digital Hospital Episode Statistics (HES) data was used to corroborate clinical data submitted by GMCs.

The av_tumour dataset was linked to genomic data on the basis of the participant identifier. Tumours labelled as either benign or in-situ were removed from the selection process, leaving only malignant, unknown or NA (the latter being the case for Genomics England participants not present in av_tumour dataset). Where av_tumour data was available for a participant it was used to confirm the tumour type submitted by the GMC. For cases where av_tumour data did not match the GMC submission, or data was not present, HES Admitted Patient Care (APC) data was used to select the closest relevant hospital appointment involving a primary diagnosis of cancer (based on ICD10 code) to the clinical sample time submitted by the GMC. If the ICD10 code for that appointment was considered a match to the tumour type submitted by the GMC, the HES data was deemed as corroborating the GMC submission.

Where HES data did not corroborate the tumour type submitted by the GMC, three additional approaches were used: i. For primary tumours, a curated set of HES operation codes was used to match between the tumour type submitted by the GMC and HES data if the operation date exactly matched the sampling date of the tumour submitted to Genomics England, ii. For non-primary tumours that were identified as colorectal by av_tumour data, and as either hepatopancreatobiliary, endometrial carcinoma or lung in the GMC submission, more flexible HES ICD10 matching was allowed provided the date difference
between the HES appointment date and tumour sampling date submitted by the GMC was <7 days, iii. For a small number of remaining samples, ICD10 and morphology data submitted by the GMC was used to corroborate the tumour type.

Tumour stage was obtained from the NCRAS dataset. Where stage_best was present in av_tumour and the date in the diagnosisdatebest column was <365 days from the clinical sample time submitted by the GMC, stage_best was used (simplified to stages 1, 2, 3 and 4) (11618/13880, 83.7%). Tumours submitted as metastatic were assigned stage 4 by default. Figo stage was used for ovarian and endometrium related clinical indications and Dukes’ stage was used for colon and rectum adenocarcinomas (both obtained from av_tumour) table.

**Survival analysis**

All survival analyses were performed in R using the survminer and survival libraries. Specifically, we used the survfit and ggsurvplot function to create the Kaplan Meier plots, and coxph for the Cox Proportional Hazard models. The ggforest function was used to create the forest plots. Date of death was obtained from Office of National Statistics (ONS) data. Where a death was not recorded for an individual, treatment and operation event dates were used to determine the last date an individual was seen in order to right-censor the data. All models were corrected by disease stage where appropriate, and group tested p-values were corrected for multiple testing using the Benjamini-Hochberg method.

**Data availability statement**

The data supporting the findings of this study are available within the National Genomics Research Library, a secure cloud workspace. To access genomic and clinical data within this Research Environment, researchers must first apply to become a member of either the Genomics England Clinical Interpretation Partnership (https://www.genomicsengland.co.uk/research/academic) or the Discovery Forum (industry partners) https://www.genomicsengland.co.uk/research/research-environment. Summary data used to generate figures in this paper are however available from the authors upon reasonable request.

**Code availability statement**

Code and data used for linking genomic data with secondary data sources are available upon request. No bespoke mathematical algorithms were used in the analysis.
Figure Legends

Figure 1. Overview of the 100,000 Genomes Cancer Programme cohort.

(A) Journey of the patient’s genome. Patients provided informed consent for paired tumour and normal (germline) whole genome sequencing (WGS) analysis. DNA was extracted from tumour and normal (blood) samples using standardised protocols and samples were submitted for whole genome sequencing, performed on an Illumina sequencer. An automated pipeline was constructed for sequence quality control, alignment, variant calling and interpretation, with results returned to the 13 NHS Genomic Medicine Centres (GMCs) for review in regional Genomic Tumour Advisory Boards. (B) Linked genomic and real-world clinical datasets. In the 100,000 Genomes Project, participants are followed over their life course with the use of electronic health records (all hospital episodes, cancer registration entries, systemic anti-cancer therapies and cause of death). (C) Distribution of 12,948 cases represented by 33 tumour types (cases with more than one sample per tumour are counted once). (D) Thirteen NHS Genomic Medicine Centres recruited patients diagnosed with cancer across England. The area of the pie chart is proportional to the number of patients recruited and the total number of participants recruited per GMC is indicated in brackets. (E) Breakdown by stage of disease (left panel), NA = not available or not applicable (in the context of glioblastoma multiforme and low grade glioma), type of sample obtained (middle panel) and tumour purity (right panel) for each tumour type, the interquartile range and median values are shown.

Figure 2. Somatic and germline alterations across common tumour types.

Prevalence of different types of mutations identified by WGS in the genes indicated for testing in the National Genomic Test Directory for Cancer (NGTDC). The first panel indicates the total percentage of cases harbouring one or more genomic alterations of clinical significance as listed in the NGTDC (where the number of cancers sequenced ≥ 10). In the subsequent panels, somatic variants (from left to right) consisting of small variants (single nucleotide variants (SNVs), insertions and deletions), copy-number aberrations (CNAs), structural variants (SVs), homologous recombination deficiency (HRD) and mismatch repair (MMR) signatures and tumour mutation burden (TMB) along with germline variants related to inherited cancer risk (predisposing genes) and pharmacogenomic (PGx) findings (toxicity-associated DYPD variants) are shown. The top five genes with the most prevalent mutation rates for each mutation type are shown (see Supplementary Figure 2 for full analysis). The percentage of tumours harbouring a specific type of mutation in gene(s) indicated for testing by tumour type in the NGTDC are shown in magenta. The incidence of mutations (as a percentage) in other tumour types, not currently indicated in the NGTDC, are shown in blue. Colour gradation indicates percentage of affected cases.
Figure 3. Predictive value of pan-genomic markers derived from WGS data

(A) Distribution of tumour mutation burden (TMB) and mutational signatures across six tumour types (samples that underwent PCR amplification during library prep are excluded and the dataset for each tumour type is downsamples to 100 samples). The horizontal red bar indicates the median TMB for each cancer type. (B) Kaplan-Meier analysis for survival (embedded forest plots indicate hazard ratios and p-values indicate cox proportional models corrected by cancer stage) in the absence and presence of the HRD signature (left panel) in cancers treated with platinum chemotherapy and in the absence and presence of the MMR signature (right panel) in cancers treated with immunotherapies. (C) Kaplan-Meier analysis for survival (embedded forest plots indicate hazard ratios and p-values indicate cox proportional models corrected by cancer stage) in skin cutaneous melanoma (left panel) and lung adenocarcinoma (right panel) in cancers with high and low TMB treated with immunotherapies. Numbers of patients at different time points are indicated on the survival curves.

Figure 4. Prognostic value of small variants and copy-number aberrations from WGS data

(A) Co-occurrence of copy-number aberrations and small variants in clinically actionable genes. Bars represent the percentage of cases with copy-number aberration in the subset of cases with or without small variants (SNV or small insertion or deletion) in clinically actionable genes. Oncogenes and tumour suppressor genes were tested for gains (red) or losses (blue), respectively, and genes indicated in red or blue show statistical difference in the percentage of cases with copy number aberrations (p-value < 0.05, Fisher’s Test) in the absence or presence of a small variant. The adjacent table indicates the level of significance for the indicated genes, including ATRX and TERT, where we examined losses and gains respectively. (B) Kaplan-Meier analysis for survival (embedded forest plots indicate hazard ratios and p-values indicate cox proportional models corrected by cancer stage) in the absence and presence of mutations in genes indicated for testing in NGTDC across all cancer types. Numbers of patients at different time points are indicated on the survival curves. Cox proportional hazard results are shown in the summary table.
Figure 2

<table>
<thead>
<tr>
<th>Tumour Type</th>
<th>Total % of Tumours with ≥1 Mutation(s) in a Target Gene(s) as Listed in the NGTDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glioblastoma Multiforme</td>
<td>95</td>
</tr>
<tr>
<td>Low Grade Glioma</td>
<td>93</td>
</tr>
<tr>
<td>Skin Cutaneous Melanoma</td>
<td>74</td>
</tr>
<tr>
<td>Head and Neck Squamous Cell Carcinoma</td>
<td>71</td>
</tr>
<tr>
<td>Colon Adenocarcinoma</td>
<td>70</td>
</tr>
<tr>
<td>Lung Adenocarcinoma</td>
<td>68</td>
</tr>
<tr>
<td>Rectum Adenocarcinoma</td>
<td>56</td>
</tr>
<tr>
<td>Ovarian High Grade Serous Carcinoma</td>
<td>55</td>
</tr>
<tr>
<td>Breast Invasive Carcinoma</td>
<td>49</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>34</td>
</tr>
<tr>
<td>Mesothelioma</td>
<td>33</td>
</tr>
<tr>
<td>Bladder Urothelial Carcinoma</td>
<td>26</td>
</tr>
<tr>
<td>Uterine Corpus Endometrial Carcinoma</td>
<td>26</td>
</tr>
<tr>
<td>Lung Squamous Cell Carcinoma</td>
<td>21</td>
</tr>
<tr>
<td>Kidney Renal Papillary Cell Carcinoma</td>
<td>16</td>
</tr>
<tr>
<td>Ovarian Endometrioid Adenocarcinoma</td>
<td>15</td>
</tr>
<tr>
<td>Esophageal Adenocarcinoma</td>
<td>12</td>
</tr>
<tr>
<td>Stomach Adenocarcinoma</td>
<td>12</td>
</tr>
<tr>
<td>Pancreatic Adenocarcinoma</td>
<td>12</td>
</tr>
<tr>
<td>Kidney Renal Clear Cell Carcinoma</td>
<td>10</td>
</tr>
<tr>
<td>Prostate Adenocarcinoma</td>
<td>5</td>
</tr>
<tr>
<td>Liver Hepatocellular Carcinoma</td>
<td>4</td>
</tr>
<tr>
<td>Ovarian Serous Carcinoma</td>
<td>4</td>
</tr>
<tr>
<td>Cholangiocarcinoma</td>
<td>3</td>
</tr>
<tr>
<td>Uterine Corpus Endometrial Serous Carcinoma</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNV/Indel</th>
<th>CNA</th>
<th>SV</th>
<th>Pan genomics markers</th>
<th>SNV/Indel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total SNV/Indel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total CNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total SV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total SV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total CNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNV/Indel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pan genomics markers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNV/Indel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3

3A

Somatic coding variants per megabase

Breast Invasive Carcinoma
Ovarian High Grade Serous Carcinoma
Uterine Corpus Endometrial Carcinoma
Colon Adenocarcinoma
Lung Adenocarcinoma
Skin Cutaneous Melanoma

3B

Key
- APOBEC
- Ageing
- Homologous recombination
- Mismatch repair
- POLE
- Smoking
- UV

3C

Skin Cutaneous Melanoma
- TMB high
- TMB low

Lung Adenocarcinoma
- TMB high
- TMB low

Survival probability vs Cancer stage

Number of cases

HR deficient (N=1625)
HR proficient (N=254)

MMR deficient (N=254)
MMR proficient (N=1625)

TMB high (N=85)
TMB low (N=48)

Number of cases

HR deficient (N=686)
HR proficient (N=1625)

MMR deficient (N=12)
MMR proficient (N=686)

TMB high (N=77)
TMB low (N=48)

Number of cases

MMR proficient (N=468)
MMR deficient (N=11)

Time (years)

0 1 2 3 4 5 6

p = 0.18

p = 0.0001

p = 0.015

p = 0.72
Figure 4

**GAIN**

- **EGFR**
- **GNAS**
- **BRAF**
- **MET**
- **KRAS**
- **ALK**
- **PIK3CA**
- **RET**
- **CTNNB1**
- **FGFR2**
- **NRAS**
- **FGFR3**
- **HRAS**
- **IDH2**
- **IDH1**

**LOSS**

- **TP53**
- **RB1**
- **CDKN2A**
- **NF1**
- **APC**
- **BRCA1**
- **ATM**
- **TSC1**
- **SMARCA4**
- **MLH1**
- **CDK12**
- **PolD1**
- **TSC2**
- **BRCA2**
- **POLE**
- **SDHA**
- **PMS2**
- **MSH2**
- **MSH6**

<table>
<thead>
<tr>
<th>Gene</th>
<th>No. of cases</th>
<th>No. of cancer types</th>
<th>p-value (corrected)</th>
<th>Significance level</th>
</tr>
</thead>
<tbody>
<tr>
<td>APC</td>
<td>2280</td>
<td>24</td>
<td>0.014</td>
<td>*</td>
</tr>
<tr>
<td>ATM</td>
<td>2840</td>
<td>31</td>
<td>0.049</td>
<td>**</td>
</tr>
<tr>
<td>ATRX</td>
<td>5356</td>
<td>31</td>
<td>0.002</td>
<td>***</td>
</tr>
<tr>
<td>BRAF</td>
<td>492</td>
<td>24</td>
<td>0.049</td>
<td>*</td>
</tr>
<tr>
<td>BRCA1</td>
<td>894</td>
<td>31</td>
<td>0.002</td>
<td>**</td>
</tr>
<tr>
<td>BRCA2</td>
<td>571</td>
<td>27</td>
<td>0.049</td>
<td>*</td>
</tr>
<tr>
<td>CDK12</td>
<td>1235</td>
<td>29</td>
<td>&lt;0.001</td>
<td>***</td>
</tr>
<tr>
<td>CDKN2A</td>
<td>1235</td>
<td>29</td>
<td>&lt;0.001</td>
<td>***</td>
</tr>
<tr>
<td>FGFR2</td>
<td>222</td>
<td>22</td>
<td>0.049</td>
<td>*</td>
</tr>
<tr>
<td>KRAS</td>
<td>1585</td>
<td>27</td>
<td>&lt;0.001</td>
<td>***</td>
</tr>
<tr>
<td>MSH2</td>
<td>792</td>
<td>29</td>
<td>0.017</td>
<td>*</td>
</tr>
<tr>
<td>MSH6</td>
<td>684</td>
<td>28</td>
<td>0.049</td>
<td>*</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>2321</td>
<td>24</td>
<td>0.017</td>
<td>*</td>
</tr>
<tr>
<td>POLD1</td>
<td>959</td>
<td>30</td>
<td>0.049</td>
<td>*</td>
</tr>
<tr>
<td>TERT</td>
<td>1154</td>
<td>26</td>
<td>0.004</td>
<td>**</td>
</tr>
<tr>
<td>TPS3</td>
<td>6968</td>
<td>31</td>
<td>&lt;0.001</td>
<td>***</td>
</tr>
</tbody>
</table>

**Survival probability**

- **Time (years)**
  - 0.00
  - 0.25
  - 0.50
  - 0.75
  - 1.00

- **Survival probability**
  - 0.00
  - 0.25
  - 0.50
  - 0.75
  - 1.00

**Number of cases**

- **Present**
  - 2280
  - 894
  - 8558
  - 1235
  - 571
  - 1235
  - 222
  - 1585
  - 792
  - 684
  - 2321
  - 959
  - 1154
  - 6968

- **Absent**
  - 2045
  - 1953
  - 1994
  - 1859
  - 1903
  - 2028
  - 2045
  - 1953
  - 1994
  - 1859
  - 1903
  - 2028

**Mutation absent**

- **Mutation present**

**Gene No. of cases No. of cancer types p-value**

- **APC**
- **BRAF**
- **CDK12**
- **CDKN2A**
- **CDK12**
- **CDKN2A**
- **FGFR2**
- **KRAS**
- **MSH2**
- **MSH6**
- **PIK3CA**
- **POLD1**
- **TERT**
- **TPS3**
Supplementary Files

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

- NMEDAN123593ManuscriptSupplFiguressubmitted.pdf