Combined burden and functional impact tests for cancer driver discovery using DriverPower

Shimin Shuai1,2*, PCAWG Drivers and Functional Interpretation Working Group, Steven Gallinger3,4, Lincoln D. Stein1,2* & PCAWG Consortium

The discovery of driver mutations is one of the key motivations for cancer genome sequencing. Here, as part of the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium, which aggregated whole genome sequencing data from 2658 cancers across 38 tumour types, we describe DriverPower, a software package that uses mutational burden and functional impact evidence to identify driver mutations in coding and non-coding sites within cancer whole genomes. Using a total of 1373 genomic features derived from public sources, DriverPower’s background mutation model explains up to 93% of the regional variance in the mutation rate across multiple tumour types. By incorporating functional impact scores, we are able to further increase the accuracy of driver discovery. Testing across a collection of 2583 cancer genomes from the PCAWG project, DriverPower identifies 217 coding and 95 non-coding driver candidates. Comparing to six published methods used by the PCAWG Drivers and Functional Interpretation Working Group, DriverPower has the highest F1 score for both coding and non-coding driver discovery. This demonstrates that DriverPower is an effective framework for computational driver discovery.

https://doi.org/10.1038/s41467-019-13929-1

1 Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada M5S 1A8. 2 Computational Biology Program, Ontario Institute for Cancer Research, Toronto, ON, Canada M5G 0A3. 3 Division of General Surgery, Toronto General Hospital, Toronto, ON, Canada M5G 2C4. 4 Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON, Canada M5G 1X5. PCAWG Drivers and Functional Interpretation Working Group authors and their affiliations appear at the end of the paper. PCAWG Consortium members and their affiliations appear online. *email: shimin.shuai@mail.utoronto.ca; lincoln.stein@gmail.com
Cancer drivers are somatic genetic alterations that confer selective advantages to tumour cells. Identification of cancer drivers is a crucial yet challenging task in cancer genomics research. There are multiple challenges. First, driver mutations generally account for only a small fraction of the somatic variations found in a typical tumour, the rest being innocent bystander ‘passenger’ mutations. Second, there is substantial intra- and inter-tumoural heterogeneity in most cancers. Both across different tumour types and across different genomic regions within the same tumour, the background mutation rate (BMR) can vary over several orders of magnitude.

The advent of large-scale cancer whole-genome sequencing (WGS) data has made it possible to explore the role of driver events in non-coding regions. However, identifying non-coding driver events in WGS creates new challenges. First, although the functional impact of somatic mutations in the coding regions of genes is fairly straightforward to predict, much less is known about the effect of mutations on non-coding regions of the genome. Second, only ~1% of somatic mutations detected in PCAWG WGS data are exonic, adding substantially more mutations and regions to be tested and demanding more careful control of type I errors in non-coding regions. However, identifying non-coding (WGS) data has made it possible to explore the role of driver elements (BMR) can vary over several orders of magnitude. 

Features predictive of BMR. To evaluate DriverPower, we took WGS somatic variant data derived from 2583 high-quality donors from the PCAWG project. After removing hypermutated samples, 2514 donors with 24,715,214 somatic single nucleotide variants (SNV) and small indels were used for driver element identification. We analysed these data both as a single pan-cancer data set, as well as a series of 29 tumour type-specific cohorts (Supplementary Data 1).

Among all tumour cohorts, we observed substantial variability in the observed mutation rate at the tissue, donor and locus levels (Supplementary Figs. 1 and 2). Accurate driver detection requires an accurate estimate of BMR across the tumour genome, taking into account the extensive variability among tumour types, donors and genomic regions. DriverPower tackles this issue by modelling the BMR using numerous genomic features that co-vary with the localised BMR. We collected 1373 features from three public data portals (Supplementary Data 2): the ROADMAP Epigenomics project, the ENCODE project and the UCSC genome browser. These features covered seven main categories: conservation, DNA accessibility, epigenomic marks, nucleotide contents, replication timing, RNA expression and genome compartments. As expected, we found extensive multicollinearity among features. Most features (1368/1373) are significantly (Spearman’s rho test q < 0.1) correlated with pan-cancer genome-wide mutation rates (Supplementary Fig. 3).

BMR model. We investigated two algorithms for modelling the BMR based on genomic features. The first algorithm was randomised lasso followed by binomial generalised linear model (GLM). The alternative algorithm was the gradient boosting machine (GBM), which is a non-linear and non-parametric tree ensemble algorithm. To evaluate both BMR modelling algorithms, we made non-overlapped 1 megabase pair (Mb) autosomal elements (n = 2521) as well as training genomic elements (n = 867,266) by sampling genomic coordinates randomly. The number of mutations per element was then predicted with five-fold cross validation (CV).

When evaluated using 1-Mb autosomal elements, we found that both algorithms could accurately predict the BMR (Supplementary Figs. 4 and 5). In high mutational burden tumour cohorts, we observed essentially no difference between two algorithms, however GBM consistently outperformed GLM when applied to low mutational burden tumour cohorts (Supplementary Fig. 6). When evaluated on the training element set, in which the size of element varies from 100 bp to 1 Mb, the prediction accuracy drops due to higher BMR variability, especially for low mutational burden tumour cohorts such as Melanoma-MPN and CNS-PiloAstro (Supplementary Fig. 6). However, for large cohort such as the pan-cancer set (N = 2253), ~93% of the mutation rate variance on the training set is explained by either model (Fig. 1b). The model still shows excellent performance when applied to the test element set, explaining 83% of the mutation rate variance on the pan-cancer cohort (Fig. 1c).

Both the randomised lasso algorithm and the GBM can be used to rank feature importance in different ways. Feature selection ranking from both methods confirmed that H3K9me3 (associated with heterochromatin), replication timing and H3K27ac (or its antagonistic histone mark H3K27me3) are the most important groups of predictors for BMR (Supplementary Fig. 7 and Supplementary Data 2). Consistent with previous results, we found that features from tumour cell lines with similar cell-of-origin to the primary tumour type are frequently selected. For example, replication timing from liver cancer cell line HepG2 was selected as a feature for the BMR in hepatocellular carcinoma (Liver-HCC), whereas replication timing in MCF7 (breast cancer) and SK-N-SH (neuroblastoma) were selected for breast adenocarcinoma (Breast-AdenoCA) and glioblastoma (CNS-GBM), respectively (Supplementary Fig. 8).
Although the DriverPower framework can potentially work with any functional scoring scheme, in the current implementation we measured the functional impact using four published scoring schemes: the CADD\textsuperscript{16}, DANN\textsuperscript{17}, EIGEN\textsuperscript{18} and LINSIGHT\textsuperscript{19} scores. Although different training data, assumptions and algorithms are used by different scores, we found those scores to be consistent at the element level (Supplementary Fig. 9). We used the average weight of all four scores in the remainder of the manuscript unless otherwise specified.

**Candidate driver event discovery.** To evaluate the DriverPower algorithm, we first employed three simulated variant sets generated by the PCAWG Drivers and Functional Interpretation Working Group (PDFIWG) to examine type I and type II errors. We expected to identify no drivers as all three simulated data sets are reshuffles of observed mutations. In general, we observed no inflation or deflation in simulations and only eight significant hits (DriverPower $q < 0.1$) were identified in $\sim 11$ M statistical tests (Supplementary Fig. 10). We then used the observed PCAWG data set to discover drivers within multiple coding and non-coding element sets identified by the PDFIWG, spanning 3.7% ($\sim 113$ Mbp) of the human genome.

We benchmarked our results against reference driver element sets and driver candidates called by six other published methods. Among the six methods, ExInAtor\textsuperscript{20}, ncdDetect\textsuperscript{21} and LARVA\textsuperscript{22} use only mutational burden information; oncodriveFML\textsuperscript{23} uses only functional biases; whereas MutSig\textsuperscript{24} and ActiveDriverWGS\textsuperscript{25} model both mutational burden and functional consequence but not through functional impact scores. Three reference driver element sets were used: the COSMIC Cancer Gene Census (CGC\textsuperscript{26,27}), the PCAWG raw integrated driver candidates (PCAWG-raw) and the PCAWG consensus driver candidates (PCAWG-consensus). The CGC is a catalogue of driver genes for which mutations have been causally implicated in cancer and was used as the gold standard set (i.e., used in the calculation of precision and recall) for coding and splice site drivers. PCAWG-raw is an integration of driver elements called by 12 different driver detection methods on the same data we used here. PCAWG-consensus is a conservative set derived from the PCAWG-raw by applying multiple stringent filters to control the false discovery rate; in particular, the majority of non-coding candidates from lymphoid tumours and skin melanomas is excluded from this set because of hyper-mutational processes in these tumour types that create prominent mutational hotspots\textsuperscript{28–30}. For the same reason our analysis of non-coding regions for tumour-specific and the pan-cancer cohorts excluded melanoma and lymphoma.

Overall, we observed well-calibrated $p$ values in DriverPower’s results with or without functional adjustment (Fig. 1d and Supplementary Fig. 10) and a high accuracy for both coding and non-coding driver discovery (Fig. 1e, Supplementary Data 3). For protein-coding regions (CDS), DriverPower found 217 significant ($q < 0.1$) driver candidates. As a gene (e.g., TP53) can be driver in multiple cohorts, the unique number of genes was 131. The precision of the algorithm’s driver calls was high. Among the driver genes called by DriverPower, 82.5% (179/217) of all genes were present within the CGC. For non-CGC genes, 27 and 3 genes were present within PCAWG-consensus and PCAWG-raw, respectively. Thus, only 3.7% (8/217) coding driver candidates...
called by DriverPower were not contained within any reference gene sets. As expected, incorporation of functional information increased both precision and recall in coding driver discovery (Fig. 2a and Supplementary Fig. 11). For example, in pancreatic ductal adenocarcinoma (Panc-AdenoCA; \( N = 232 \)), the addition of functional adjustment to the algorithm resulted in a gain of three additional drivers (ACVR1B, RBM10 and ZFP36L2) and the loss of one likely false-positive genes (FAU) (Fig. 2a). Without the use of functional information, the overall precision dropped to 74.6% (156/209) for CGC genes and 88.5% (185/209) for CGC/non-melanoma/lymphoma. DriverPower without mutated element. Copy number corrected of 26 tumour cohorts (no melanoma and lymphoma).

**Fig. 2 Benchmarking DriverPower driver discovery performance.** a Comparison of CDS results with or without functional adjustment for Panc-AdenoCA. Dashed lines in a represent the q value = 0.1 threshold. Function-adapted q values are q values with the incorporation of functional impact scores. Only significant genes are labelled (colour legend is the same as Fig. 1e). b, c Benchmark results for coding genes compared with six other driver discovery methods. d, e Benchmark results for 3′-UTR, 5′-UTR, promoter and enhancer sets. b, d Show the precision and recall for each method according to results of 26 tumour cohorts (no melanoma and lymphoma). e Shows the number and fraction of coding driver candidates called by each method that are contained within reference gene sets. The coloured columns in e correspond to different reference driver sets (colour legend is the same as Fig. 1e). e Shows the number of hits and fraction of non-coding driver candidates called by each method that are also called by others. The coloured columns in e correspond to the number of methods that agree on a driver candidate. f Differential expression analysis for the CDS and splice site of SGK1 in Lymph-BNHL. g Differential expression analysis for the GPR126 enhancer in Bladder-TCC. MUT indicates samples with mutated element and WT indicates samples without mutated element. Copy number corrected \( p \) values from the likelihood ratio test and the log2 fold changes (log2FC) are shown in blue.

We next benchmarked DriverPower’s accuracy for non-coding driver events. For the prediction of driver events affecting the splice sites of coding genes, DriverPower called 47 significant (\( q < 0.1 \)) candidates with 85.1% (40/47) within CGC. DriverPower (F1 = 0.91) also outperformed two recently published methods, ncdDetect (F1 = 0.65) and oncoDriverFML (F1 = 0.32), for splice site driver detection (Supplementary Fig. 13). For the prediction of non-coding driver events in 3′-UTRs, 5′-UTRs, promoters and enhancers, DriverPower identified 19 candidates in non-melanoma/lymphoma tumour cohorts and 24 candidates in the pan-cancer cohort. Benchmarking results showed that DriverPower has the highest F1 score (0.79) among the six methods evaluated (Fig. 2d, e). Promoter and 5′-UTR driver candidates called by DriverPower were associated with a total of 17 unique genes. Of these, one gene (TERT) was in CGC, four genes (WDR74, HESI, MTG2 and PTDSII) were in PCAWG-consensus, and six other genes were in PCAWG-raw. DriverPower also called two 3′-UTR driver candidates in total, including TOBI in pan-cancer and ALB in Liver-HCC. Both candidates were present in the PCAWG-consensus. For enhancer regions, DriverPower identified two candidates: chr6:142,705,600-142,706,400 (linked to GPR126) and chr7:86,865,600-86,866,400 (linked to TP53TG1). Both enhancer elements were identified by PCAWG-raw; the TP53TG1 enhancer was the only enhancer for non-melanoma/lymphoma tumours in the PCAWG-consensus set.
For long non-coding RNA (lncRNA) genes and their promoters, DriverPower found 9 candidates in total. Among them, 6 and 3 were contained within PCAWG-consensus and PCAWG-raw, respectively. These candidates targeted three unique lncRNAs: RN7SK, RMRP and RPPH1. The promoter of RMRP was significantly ($q < 0.1$) mutated in four cohorts (Breast-AdenoCA, Liver-HCC, Stomach-AdenoCA and pan-cancer) and has been nominated as a novel non-coding driver.

**DriverPower-exclusive driver candidates overview.** A total of 11 coding and 17 unique non-coding candidates were exclusively identified by DriverPower (not present in either CGC or PCAWG-consensus; Supplementary Data 4). We sought to evaluate these exclusive driver candidates using literature evidence and correlative orthogonal data such as the effect of the variant on RNA-seq expression levels and the presence of somatic copy number alterations (SCNAs) and somatic structural variations (SVs) covering the same regions. On this basis, we found that many of the DriverPower-exclusive candidates are plausible cancer drivers.

Among protein-coding genes, DriverPower identified EEF1A2 (eukaryotic translation elongation factor 1 alpha 2) in the oesophageal adenocarcinoma cohort (Eso-AdenoCA; 7/95 samples). All seven observed mutations were missense (Supplementary Fig. 14a). Although no RNA-seq data are available for Eso-AdenoCA samples, SCNA analysis indicated that EEF1A2 is amplified in 69.5% (66/95) of Eso-AdenoCA samples (vs. 27.9% of non-Eso-AdenoCA samples; Supplementary Fig. 14b), suggesting a potential gain-of-function role in this cancer type. The amplification of EEF1A2 (20q13.33) was also confirmed by the GISTIC2.0 ($q = 0.0006$). The same 1-Mbp locus detected by GISTIC2.0 was also amplified recurrently in other tumour types, including 73.1% of colorectal adenocarcinoma, 64.7% of stomach adenocarcinoma and 55.4% of ovarian adenocarcinoma. Supporting this hypothesis, previous studies have also demonstrated that EEF1A2 is a putative oncogene in ovarian cancer and overexpressed in various tumour types.

Another protein-coding gene exclusively identified by DriverPower was MEF2B in B-cell non-Hodgkin’s leukaemia (Lymph-BNHL; 8/105 samples). Among nine observed mutations, eight mutations were missense and one was a frameshift deletion (Supplementary Fig. 14c). RNA-seq data confirmed that mutated samples overexpressed MEF2B (copy number corrected $p = 0.011$; Supplementary Fig. 14d). MEF2B (Myocyte enhancer factor 2B) has been identified in multiple WES studies and a previous study has also shown that MEF2B mutations can dysregulate cell migration in non-Hodgkin lymphoma.

One splice site candidate exclusively called by DriverPower is SGKI (serum/glucocorticoid regulated kinase 1) in Lymph-BNHL. The same gene was also significant in DriverPower’s CDS result for Lymph-BNHL (Supplementary Fig. 14e), resulting in a total of 13.3% (14/105) Lymph-BNHL samples being affected by non-synonymous or splice site mutations in SGKI. SGKI is present in PCAWG-raw but was filtered out owing to the large number of AID-related variants in this tumour cohort. However, differential expression analysis indicated that SGKI was significantly overexpressed in mutated Lymph-BNHL samples relative to non-mutated samples (copy number corrected $p = 3e-13$ from likelihood ratio test; Fig. 2f). SGKI encodes a serine/threonine protein kinase that has an important role in cellular stress response and its CDS has been nominated as a driver in earlier WES studies. Another study has also demonstrated that the administration of an SGKI inhibitor induces apoptosis in lymphoma cell lines. Together these data support a potential driver role for SGKI in Lymph-BNHL.

The GPR126 (adhesion G protein-coupled receptor G6) enhancer candidate was filtered out from the PCAWG-raw set because of mutations in palindrome loops, which makes it unclear whether mutations in the GPR126 enhancer are caused by mutational mechanism associated with palindrome loops or positive selection. We found that the GPR126 enhancer is recurrently mutated in transitional cell carcinoma of the bladder (Bladder-TCC; 14/23 samples) and breast adenocarcinoma (Breast-AdenoCA; 8/195) (Supplementary Fig. 14f). GPR126 is among the MammaPrint 70 gene panel used to predict the risk of breast cancer metastasis. A study also shows that knockdown of GPR126 can inhibit the hypoxia-induced angiogenesis in model organisms. Differential expression analysis demonstrated that the GPR126 is significantly downregulated in Bladder-TCC samples with enhancer mutations (copy number corrected $p = 0.012$ from likelihood ratio test; Fig. 2g) relative to those carrying the wild-type enhancer, suggesting a functional role for these mutations.

Several somatically altered histone genes have been implicated in human cancer, such as H3F3A (identified as a pan-cancer driver in this study), H3F3B and HIST1H3B. DriverPower identified four histone genes as driver candidates in the pan-cancer cohort, two of which were absent from CGC or PCAWG-consensus: the 5′-UTR of HIST1H2AC and HIST1H2BD. Previous studies have shown that the protein levels of the replication-dependent histone H2A variant HIST1H2AC (encoding histone H2A type 1-C) is decreased in chronic lymphocytic leukaemia patients and bladder cancer cells and the siRNA knockdown of HIST1H2AC increases cell proliferation and promote oncogenesis.

Several other driver candidates exclusively called by DriverPower are associated with genes that may have a role in cancer. The highly expressed liver-specific gene ALB (albumin) is significant (DriverPower $q < 0.1$) for somatic mutations affecting its CDS, splice site, 3′-UTR and promoter in Liver-HCC; the splice site and promoter (under CADD scores) were discovered by DriverPower exclusively. Correlative evidence from gene expression and copy number alterations suggested that loss-of-function mutations in ALB are subject to positive selection in Liver-HCC as described elsewhere. The CDS of KAT8 (lysine acetyltransferase 8) was called by DriverPower in Panc-AdenoCA with 100% (5/5) missense mutations. As a histone modifier, KAT8 has been shown to physically interact with MLL1 and regulate known cancer drivers ATM and TP53. Previous studies have also shown that KAT8 is downregulated in gastric cancer and that KAT8 can suppress tumour progression by inhibiting epithelial-to-mesenchymal transition. The 5′-UTR and promoter of SRSF9 (serine and arginine rich splicing factor 9) was significant in DriverPower’s results for pan-cancer and not present in any reference driver sets. The protein encoded by SRSF9 is part of the spliceosome; a previous study indicates that the proto-oncogene SRSF9 is overexpressed in multiple tumours and that this overexpression can cause the accumulation of β-Catenin. The same study also showed that the depletion of SRSF9 proteins could inhibit colon cancer cell proliferation.

In summary, 4/11 coding and 4/17 unique non-coding driver candidates exclusively called by DriverPower had some form of support from the literature or orthogonal evidence. If we assume that all the exclusive candidates that lack such evidence are false positives, then this puts an estimate of DriverPower’s false discovery rate across the PCAWG data set at 3.2% (7/217) for coding and 16.8% (16/95) for non-coding regions. However, this assumption is probably invalid as most of these lack-of-evidence candidates are also identified by other methods and present in PCAWG-raw. We acknowledge that lack-of-evidence candidates may contain false-positive calls, but they may also contain previously unknown drivers. For example, the 5′-UTR of
**TBC1D12** in Breast-AdenoCA, which has been filtered out from the PCAWG-raw owing to possible hypermutability, is called by all but one driver discovery methods and is reported as a putative cancer driver in previous studies because of two recurrent mutations in the Kozak consensus sequence involving in the initiation of protein translation23,35. Moreover, according to another recent study, the same TBC1D12 candidate is still statistically significant in breast cancer even after removing hypermutations, but whether these mutations can alter protein translation in cancer is still undetermined24. Some lack-of-evidence candidates may also fit the mini-driver model of cancer evolution56. Unlike classical drivers, mini-drivers can only weakly promote and are not essential for tumour progression, hence present at a lower frequency in cancer cohorts. Further investigation is required to determine the role of lack-of-evidence candidates in cancer.

**DriverPower applied to WGS.** To demonstrate the robustness of DriverPower, we applied DriverPower to two public whole-exome sequencing (WES) data sets (Supplementary Fig. 15). Both WES data sets are processed differently than the PCAWG data and contain samples not included in the PCAWG study. For liver cancer, using models trained for Liver-HCC (N = 314), DriverPower identified 14 coding drivers from 364 TCGA-LIHC samples (53 shared with Liver-HCC). All but one driver candidates were present within the CGC or PCAWG-consensus. For pancreatic adenocarcinoma, using models trained for Pan-AdenoCA (N = 232), DriverPower identified six coding drivers from 180 TCGA-PAAD samples (no shared samples with the PCAWG study) and all corresponded to known driver genes.

**Discussion**

Computational driver discovery is essential to distinguish driver from passenger mutations in the coding and non-coding regions of whole cancer genomes. Here we report DriverPower, a new framework for accurately identifying both types of driver mutation by combining mutational burden and functional impact information. The method takes advantage of the large somatic mutation sets produced by WGS technology to build an accurate global BMR model from more than a thousand genomic features. This contrasts with methods that build a local BMR model using selected or flanking regions. One advantage of this is that the method is not biased towards coding regions, but uses the same model for coding and non-coding cancer driver discovery. Another advantage is the method’s high degree of modularity. DriverPower can potentially work with any types of genomic element (contiguous or disjoint, coding or non-coding, proximate or distal to genes), any regression algorithms for modelling BMR and any functional impact score scheme. Although DriverPower is designed for WGS projects, it performs robustly in WES strategies as well.

In comparison with the other driver discovery methods evaluated by the PCAWG Drivers and Functional Interpretation Working Group, DriverPower provides the best balance of precision and recall, although it is not always the top-ranked method when either metric is considered independently (Fig. 2b, d). As discussed in Supplementary Note 1, DriverPower is parameterised to allow for adjustment of the precision-recall trade-off; in this study, we selected conservative parameters that prioritise precision over recall especially for non-coding regions (Supplementary Fig. 16).

There are several ways in which the accuracy of DriverPower could be improved. One approach to improve recall is to take into the account the potential presence of negative (purifying) selection in the functional regions selected for testing. When the BMR model is trained, we use random genomic elements that are predominantly under neutral selection. However, the functional elements selected for testing are more likely to be under positive and/or negative selection47. The observed mutation rate reflects the balance between positive and negative selection, and negative selection at one site in the element will diminish the signal of positive selection at other sites, reducing the sensitivity of the method as a whole. To our knowledge, no driver discovery tool currently models the effect of negatively selected sites; future work aims to take this mechanism into account.

The precision of the method can also be improved. False-positive driver calls may be caused by technical errors such as variant-calling artefacts that artificially increase the local mutation rate, or by biological processes that are not captured by the BMR model such as regional differences in activation-induced cytidine deaminase (AID) activity. These can potentially be mitigated by incorporating into the BMR model additional features relevant to the technical and biological processes. For example, incorporating read-level coverage, mapping and bias scores into the BMR could help correct for regions prone to variant-calling artefacts, whereas features like the number of palindrome loops and the fraction of mutations caused by AID per element could be used to adjust for locally-acting hypermutation processes.

When applied to the PCAWG data set, DriverPower called nearly twice as many coding driver events as non-coding ones, a ratio also observed by the PCAWG driver study, a ratio also observed by the PCAWG driver study47. Although this unbalanced ratio may reflect cancer biology, there is also the possibility that it reflects, at least in part, the technical challenge of sequencing and interpreting non-coding regions. Potential artefacts include systematic undercalling of somatic variants in non-coding regions24, a problem that could be rectified by deeper coverage. For example, it is estimated that ~216 mutations in the TERT promoter are likely to be missed in the PCAWG data set owing to low detection sensitivity47. Another technical issue is raised by the fact that several non-coding candidates are only significant in the pan-cancer cohort, suggesting that the data set is statistically underpowered. In fact, although we studied 2583 genomes here, many tumour types have a sample size fewer than 30. To overcome this issue, we could either sequence more genomes or reduce the size of the set of test elements by narrowing it to functional motifs or conserved bases58. Moreover, only ~3.7% of the genome has been studied here. There may be more non-coding drivers in other types of regulatory elements, which demands more complete annotations for the non-coding part of the human genome. At last, functional impact score schemes are currently biased toward coding mutations; therefore, improved functional scoring schemes will also help us identify more functionally relevant non-coding cancer drivers in the future.

A comprehensive catalogue of coding and non-coding cancer drivers will accelerate the clinical translation of cancer genomic study to precision medicine. As more cancer genomes and more cancer types are sequenced, a general and accurate framework for computational driver discovery like DriverPower will become increasingly useful.

**Methods**

**Ethical review.** Sequencing of human subjects tissue was performed by ICGC and TCGA consortium members under a series of locally approved Institutional Review Board (IRB) protocols as described in Hudson et al.19. Informed consent was obtained from all human participants. Ethical review of the current data analysis project was granted by the University of Toronto Research Ethics Board (REB) under protocol #30278, ‘Pan-cancer Analysis of Whole Genomes: PCAWG.’

**Generation of cancer whole-genome somatic mutations.** All DNA somatic SNVs and indels for 2583 donors were obtained from the PCAWG project (somatic variant callset released October 2016)18. For our analysis, donors with hypermutated signatures were excluded (n = 69, defined as >30 mutations per Mb). Otherwise,
we used the same type-specific \( (n = 29) \) and pan-cancer (all tumour samples except Skin-Melanoma, Lymph-NOS, Lymph-CLL and Lymph-BNHL) sample cohorts as the PCDWG Drivers and Functional Interpretation Working Group (PDFIWG; Supplementary Data 1)\(^7\).

**Generation of simulated somatic mutations.** We used three simulated data sets (Broad, DKFZ and Sanger simulations) from the PDFIWG (described in detail at Rheinbay et al.\(^47\)). These simulations were made to capture the variation of BMR and remove the signal of positive selection through permutations of observed somatic mutations.

**Generation of test and training genomic elements.** We define a genomic element as the collection of genome coordinates that defines one specific functional region of interest. For example, the CDS element of TP53 is the combination of all protein-coding regions in TP53.

We used eight test element sets in our analysis, including the CDS \((n = 20,185)\), splice site \((n = 18,729)\), 5′-UTR \((n = 19,369)\), 3′-UTR \((n = 19,188)\), enhancer \((n = 30,816)\), IncRNA \((n = 5,580)\) and IncRNA promoter \((n = 53,737)\). All test element sets were obtained from the PCAWG project. GENCODE v19 was used as the reference gene model when building those sets\(^58\). Non-coding RNA annotations were collected from multiple sources as described.

We constructed genomic element training sets by randomly sampling genome coordinates from hg19, the build used for PCAWG. The length of each training element was sampled from the length distribution of test elements and multiplied by a factor of 3. Training elements overlapping test elements were removed. In total, 82266 training elements were created and ~54% \((1,545,491,987\) bp) of the genome was covered.

**Collection and generation of features.** We collected 1373 features in total. Details including data sources can be found at Supplementary Data 2. Nucleotide content features were calculated as the fraction of 2-mers and 3-mers in each genomic element. The number of 2-mers and 3-mers was counted directly from genome sequences (hg19). For raw features in bigwig format (typically genome-wide signals), we calculated the average signal strength of covered bases in each element, using the bigWigAverageOverBed (v2) utility from the UCSC genome browser\(^44\). For raw features in BED format (typically narrow peaks of ChIP-seq data), we calculated the percentage of bases intersecting BED for each element with the BEDTools (v2.24.0)\(^55\). All missing values in features were filled with 0.

**The DriverPower outline.** The main steps of the DriverPower (v1.0.0) framework are summarised below. Details of each step are described in following sections. The difference between v1.0.0 and the version used in the PDFIWG data analysis freeze (April 2017) is discussed in Supplementary Note 2 and Supplementary Fig. 17.

The first step of DriverPower is to scale features and/or filter out excluded regions. The second step is to build the BMR model using the GBM, or randomised lasso followed by binomial GLMs. The purpose of the BMR model is to estimate the expected number of mutations \((y^\omega)\) for any genomic element. Namely, we want to obtain \(y^\omega = E[y|X, L]\) where \(X\) and \(L\) are the feature vector and length for the element \(i\). The third step is to conduct burden test with observed \(y\) and predicted \((y^\omega)\) mutation counts, and perform multiple testing correction. The fourth step is to adjust observed mutation counts \(y\) based on functional impact scores for nearly significant elements \((q < 0.25)\). The last step is to re-assess the significance for nearly significant elements with functional adjusted mutation counts followed by multiple testing correction.

**Scaling of features.** Features were scaled with RobustScaler from scikit-learn (version 0.18)\(^63\). Feature scaling was only conducted for randomised lasso and GLMs.

**Definition of excluded regions.** In this study, all bases in the excluded regions were removed before any analysis. The excluded region consists of three sets: (1) all \(N\) bases and gaps in the hg19 genome (fetched from the UCSC table browser\(^12\)); (2) ENCODE consensus excluded regions (the DAC blacklisted Regions track and the Duke Excluded Regions track from the UCSC genome browser)\(^64\); (3) PCAWG low mappability regions (data retrieved from the PCAWG variant group). PCAWG low mappability regions are defined as regions callable in fewer than 556/1111 (~ 50%) tumour-normal pairs. For each tumour-normal pair, a base is callable if there are more than 14/8 high quality reads in tumour/normal WGS. In total, 2,806,377,226 bp, or 96.41% of the genome are defined as callable.

**Feature selection with randomised lasso.** To select features, we randomly sub-sampled 10\% of the training set 500 times. Then for the \(k\)-th subset with size \(N_k\), the following model was fitted\(^59\):

\[
\hat{w}^*_k = \arg \min \left( \frac{1}{N_k} \logit \left( \frac{y^\omega + 1/2}{N - L} \right) - Xw \right)^2 + a \sum_{i=1}^{137} \frac{|w_i|}{b_i} \]

where \(N\) is the total number of donors in the data set, \(X\) is the feature matrix, \(w\) is the weight vector, \(a\) is the regularisation parameter, and \(b\) is the scaling factor. The regularisation parameter \(a\) was determined by a fivefold CV lasso with 33% of the training data. For the \(k\)-th sub-sampling, the \(i\)-th feature was selected if \(w_i > 0.001\). The final feature importance score was calculated as the fraction of times that a feature was selected. Only features with score > 0.5 were used in the GLM BMR model.

**Prediction of the BMR with GLM.** When using the GLM, we modelled the observed mutations in each genomic element with a binomial distribution, that is \(y \sim B(n, p)\) with \(n = N \cdot L\) and \(p = \gamma^\omega / (N - L)\)

\[
y = \frac{\gamma^\omega}{N - L} = \logit^{-1}(\chi_{L}^{\omega} \beta)
\]

where \(\chi_{L}^{\omega}\) is the selected feature matrix and \(\beta\) is the regression coefficient vector.

**Prediction of the BMR with GBM.** We trained the GBM with XGBoost\(^66\). All features were used in model training. The negative Poisson log-likelihood was chosen as the objective function and \(\text{bin}(N - L)\) of elements were used as offset (i.e., base_margins in XGBoost). Other non-default parameters used in DriverPower were as follows: \(eta = 0.05\), \(max_depth = 8\), \(subsample = 0.6\), \(max_delta_step = 1.2\), \(early_stop_rounds = 5\) and \(nrounds = 5000\). The feature importance for GBM is measured by the improvement in accuracy brought by a feature across all trees. XGBoost returns feature importance that sums up to 1 for all features. We also normalised the importance to a \([0, 1]\) scale (i.e., importance relative to the most important feature).

**Evaluation of two BMR models.** We evaluated both models with 1 Mb autosomal bins \((n = 2521)\) and training genomic elements \((n = 867,266)\) defined above. The 1 Mb elements have been used in many studies\(^44,45\). For both elements, we obtain the predicted mutation rate by fivefold CV. For 1 Mb elements, we used fourfold data for model training and onefold data for model evaluation. For training elements, we use 1-fold data to train the model and the rest to evaluate. As per previous work, we used \(R^2\) score and Pearson’s \(r\) as evaluation metrics\(^15\). Standard error of the mean (SEM) for \(R^2\) and \(r\) was calculated from fivefold CV.

**Calculation of element functional impact scores.** Four different functional scores were used in this analysis\(^68-71\). For CDS, CADD (SNVs and indels, v1.3), DANN (SNVs) and EIGEN (SNVs) scores were used. CADD indel scores were generated with the CADD web interface for all observed indels in the PCAWG data set. For splice site, CADD and DANN scores were used. For non-coding elements, the CADD, DANN and LINSIGHT (SNVs and indels) score were used. Then the following steps were used to calculate the functional impact score per genomic element. First, raw scores were retrieved for all observed mutations in the data set. Second, all raw scores were converted to p-value-like scores by \(-10 \log_{10}(\text{rank}/N)\), where \(N\) is the number of observed mutations having scores. Third, for each genomic element, its functional score \(S\) was calculated as:

\[
S = \frac{1}{N} \sum_{i=1}^{N} S_i
\]

where \(N\) is the number of donors and \(S_i\) is the average functional impact score for the \(i\)-th donor.

**Adjustment of the mutation count.** To compensate for the unbalanced number of mutations among samples, instead of using the mutation count per element directly we used the geometric mean of mutation count and sample count. That is, we use the balanced count \(y^\omega = \sqrt{N} y^\omega\) instead of \(y^\omega\) directly for significance test, where \(N\) is the number of mutated donors. Based on the motivation that not all mutations should be weighted the same, the balanced mutation count \(y^\omega\) was then adjusted for nearly significant elements (raw \(q\) value < 0.25) by a functional weight \(w\), that is \(\hat{y} = a n, \hat{y}^{\omega}\) where \(\hat{y}^{\omega}\) is the functionally adjusted mutation count. For the element \(j\), the functional weight \(w_j\) was calculated based on its functional score \(S_j\) and a threshold score \(S_T\):

\[
w_j = S_j / 10 \log_{10} R^2
\]

The threshold score \(S_T\) is controlled by a single parameter \(F\) between 0 and 1, and can be interpreted as the fraction of functionally relevant variants among all observed variants. Parameter tuning of \(F\) can be found at Supplementary Note 1.

**Assessment of the element significance.** For each element, we calculated \(P(\hat{y}^\omega > y^{\omega})\) as the raw \(p\) value and \(P(\hat{y}^\omega > y^{\omega})\) as the function-adapted \(p\) value. As over-dispersion has been documented in burden-based methods and can affect the driver discovery accuracy\(^22\), here we performed a regression-based overdispersion test for each tumour cohort using the training set\(^68\). Based on the result of the
overdispersion test, we calculated the raw and function-adopted p values by following a binomial distribution or a negative binomial distribution:

\[ y' \sim \begin{cases} \text{NB}(y', \theta) & \text{if } p \leq 0.01 B \left( N \cdot L \cdot y', N-L \right) \\ \text{NB}(y', \theta) & \text{otherwise} \end{cases} \]

otherwise, where \( p \) and \( \theta \) are the p value and dispersion parameter estimated from the overdispersion test, and \( s \) is the scaling factor for \( \theta \) used to accommodate the discrepancy between test and training set in terms of the dispersion level. We used \( s = 3 \) for lymphomas and \( s = 1 \) otherwise in this analysis.

**Multiple testing correction.** In all cases, \( q \) values were generated by the Benjamini–Hochberg procedure. We chose \( q < 0.1 \) as the significant level and \( q < 0.25 \) as the nearly significant level. For each element set, multiple testing correction was performed for each tumour cohort (cohort q value) and across all tumour cohorts (global q value). Cohort q values were used in functional adjustment and global q values were used to define the final driver list.

**Generation of reference cancer drivers.** Reference cancer drivers were used to benchmark the performance of DriverPower. Three reference sets were used: (1) the COSMIC CGC (v82, \( n = 567 \)); (2) the PCAWG consensus driver-candidates (PCAWG-consensus; \( n = 157 \) for coding and \( n = 26 \) for non-coding); (3) the PCAWG raw integrated driver candidates (PCAWG-raw; \( n = 193 \) for coding and \( n = 79 \) for non-coding). PCAWG-consensus (q value post-filtering <0.1) is a set of highly confident non-coding drivers and subjected to multiple stringent filters as described. PCAWG-raw (q value pre-filtering <0.1) is a superset of the consensus and includes non-coding drivers that were not subjected to the filtering process. PCAWG-raw driver candidates that are mutated in fewer than three samples were removed in this analysis. For promoter and 5'–UTR candidates in the PCAWG consensus drivers, we reversed the filtering for overlapping elements (i.e., one element is selected over the overlapping element based on prior knowledge). For example, we kept both the promoter and the overlapping 5'–UTR of WDR74 in this analysis; in the PCAWG consensus set, the WDR74 promoter is preferentially selected over its 5'–UTR.

**Benchmarking of DriverPower.** We compared coding and non-coding driver candidates called by DriverPower to driver candidates called by six other published driver detection methods (ActiveDriverWGS, ExInAtor, LARVA, MutSig2, ncddetect and oncodriveFML). Driver calls for 26 single tumour cohorts (no Skin-Melanoma, Lymph-CLL and Lymph-BNHL) were retrieved from the PCAWG driver group. For each method, we removed driver candidates that are mutated in fewer than three samples. We used precision (TP/(TP + FP)), recall (TP/(TP + FN)) and F1 score (2*Precision*Recall/(Precision+Recall)) as performance metrics.

**WES data analysis.** We obtained two WES data sets through theGenomic Data Common (GDC)78: TCGA-PAAD (35,321 somatic mutations across 180 samples) and TCGA-LIHA (56,208 somatic mutations across 364 samples). We chose public MuTect variants from GDC. Variant coordinates were lifted from hg38 to hg19 with the UCSCliftOver tool. Only CADD scores were used to detect drivers. For TCGA-PAAD, GBM models trained from Panc-AdenoCA of the PCAWG data were used. For TCGA-LIHA, GBM models trained from Liver-HCC were used.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Code availability.** The source code for DriverPower (written mainly in Python 3) is available at GitHub [https://github.com/smshuai/DriverPower]. It is distributed under GNU General Public License v3.0, which allows for reuse and redistribution. Other software packages and bioinformatics tools used in this study are indicated in the corresponding method sections. The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at [https://dockstore.org/search?search=pcawg] under the GNU General Public License v3.0, which allows for reuse and distribution.

**Data availability.** The data sets underpinning the analyses in the paper are detailed in Supplementary Table 1. Aligned sequencing data, as well as somatic and germline variant calls from PCAWG tumours, including SNVs, indels, copy number alterations and structural variants, are available for download at [https://dcc.icgc.org/releases/PCAWG]. Additional information on accessing the data, including raw read files, can be found at [https://docs.icgc.org/pcawg/data/]. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in open tier, which does not require access approval. To access potentially identification information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP [https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login] for access to the TCGA portion of the data set, and to the ICGC Data Access Compliance Office (DACO) [https://icgc.org/daco] for the ICGC portion. In addition, to access somatic SNVs derived from TCGA donors, researchers will also need to obtain dbGaP authorisation.

In addition, the analyses in this paper used several data sets that were derived from the raw sequencing data and variant calls (Supplementary Table 1). The individual data sets are available at Synapse ([https://www.synapse.org/]), and are denoted with synXXXXX accession numbers (listed under Synapse ID); all these data sets are also mirrored at [https://dcc.icgc.org], with full links, filenames, accession numbers and descriptions detailed in Supplementary Table 1. The data sets encompass: harmonised tumour histopathology annotations using a standardised hierarchical ontology (syn10389164); consensus somatic SNVs, MNVs and indels (syn7364923); gene expression profiles from RNA-sequencing data (syn5553991); genomic intervals used in driver region calls (syn5259890); driver region calls by each individual methods (syn7359546); consensus gene-level somatic copy number calls (syn239175); three simulation data sets for somatic mutations (syn1787923, syn7436065 and syn1752699).

Received: 21 November 2017; Accepted: 9 December 2019; Published online: 05 February 2020

**References**

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
2. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
3. Kikumori, E. et al. Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **17**, 93–108 (2016).
4. Garraway, L. A. & Lander, E. S. Lessons from the cancer Genome. *Cell* **153**, 17–37 (2013).
5. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
6. Lavrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
7. Mansour, M. R. et al. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377 (2014).
8. Huang, F. W. et al. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
9. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* https://doi.org/10.1038/s41586-020-1969-6 (2020).
Acknowledgements

This work was supported in part by the Government of Ontario. We acknowledge the contributions of the many clinical networks across ICGC and TCGA who provided samples and data to the PCAWG Consortium, and the contributions of the Technical Working Group and the Germline Working Group of the PCAWG Consortium for curation, realignment and harmonised variant calling of the cancer genomes used in this study. We thank the patients and their families for their participation in the individual ICGC and TCGA projects.

Author contributions

S.S. and L.S. conceived the study and wrote the paper. S.S. designed the software and conducted the data analysis. L.S. and S.G. supervised the project. Input for driver region calls and driver candidates called by other methods were made available by the PCAWG Drivers and Functional Interpretation Working Group, where Mark Gerstein, Gad Getz, Michael S. Lawrence, Jakob Skou Pedersen, Benjamin J. Raphael, Joshua M. Stuart and David A. Wheeler were working group co-leaders.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-13929-1.

Correspondence and requests for materials should be addressed to S.S. or L.D.S.

Peer review information Nature Communications thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PCAWG Drivers and Functional Interpretation Working Group

Federico Abascal5, Samirkumar B. Amin6,7,8, Gary D. Bader1, Prattiti Bandopadhayay9,10,11, Jonathan Barenboim2, Rameen Beroukhim9,12,13, Johanna Bertl14,15, Keith A. Boreovich16,17, Søren Brunak18,19, Peter J. Campbell5,20, Joana Carlevaro-Fita21,22,23, Dimple Chakravarty24,25, Calvin Wing Yiu Chan26,27, Ken Chen28, Jung Kyoong Choi29, Jordi Deu-Pons30,31, Priyanka Dhingra32,33, Klev Diamanti34, Lars Feuerbach35, J. Lynn Fink36,37, Nuno A. Fonseca38,39, Joan Frigola30, Carlo Gambacorti-Passerini40, Dale W. Garsed41,42, Mark Gerstein43,44,45, Gad Getz9,13,46,47, Qianyun Guo48, Ivo G. Gut49,50, David Haan51, Mark P. Hamilton52, Nicholas J. Haradhval9,53, Arif O. Harmanci45,54, Mohamed Helmy55, Carl Herrmann26,56,57, Julian M. Hess58,59, Asger Hobolth14,48, Ermin Hodzic59, Chen Hong27,35, Henrik Hornshej15, Keren Isaev2,60, Jose M.G. Izarzugaza18, Rolly Johnson21,22, Todd A. Johnson16, Malene Juul15, Randi Istrup Juul15, Andre Kahles61,62,63,64,65, Abdullah Kahraman66,67,68, Manolis Kellis9,69, Ekta Khurana32,33,34,35,36, Jaegil Kim9, Jong K. Kim72, Youngwook Kim73,74, Jan Komorowski34,75, Jan O. Korbel39,76, Sushant Kumar44,45, Andrs Lanzos21,22,23, Erik Larsson77, Michael S. Lawrence9,16,53, Donghoon Lee45, Kjong-Van Lehmann61,63,64,65,78, Shantao Li45, Xiaotong Li45, Ziao Lin9,79, Eric Minwei Liu32,33,80, Lucas Lochofsky,8,44,45, Shaoke Lou44,45, Tobias Madsen15, Kathleen Marchal81,82, Iñigo Martincorena5, Alexander Martinez-Fundichely32,33,71, Yosef E. Maruvka9,53,58, Patrick D. McGillivray44, William Meyerson45,83, Ferran Muños31,84, Loris Mularoni31,84, Hideaki Nakagawa17, Morten Muhlig Nielsen15, Marta Paczkowska2, Keunchil Park85,86,87, Jakob Skou Pedersen15,48, Tirso Pons88, Sergio Pulido-Tamayo81,82, Benjamin J. Raphael89, Jüri Reimand2,60, Iker Reyes-Salazar84, Matthew A. Reyna89, Esther Rheinbay9,13,53, Mark A. Rubin21,90,91,92,93, Carlota Rubio-Perez31,84,94, S. Cenk Sahinalp59,95,96, Gordon Saksena9, Leonidas Salichos44,45, Chris Sander61,97,98, Steven E. Schumacher9,99, Mark Shackleton41,42, Ofer Shapira9,99, Ciuye Shen9,100, Raunak Shrestha9, Shirmin Shuai1, Nikos Sidiropoulos101,102, Laura Sieverling27,35, Nasa Sinnot-Armstrong9,102, Lincoln D. Stein1,2, Joshua M. Stuart51, David Tamborero31,84, Grace Tiao9, Tatsuhiko Tsunoda16,103,104,105, Husen M. Umer34,106,
