Chromosomes and Gene Expression

Transcriptional immune suppression and upregulation of double stranded DNA damage and repair repertoires in ecDNA-containing tumors

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
February 27, 2024 (this version)

Reviewed preprint version 1
August 10, 2023

Sent for peer review
June 13, 2023

Posted to preprint server
April 24, 2023

Miin S. Lin, Se-Young Jo, Jens Luebeck, Howard Y. Chang, Sihan Wu, Paul S. Mischel, Vineet Bafna

Bioinformatics and Systems Biology Graduate Program, University of California at San Diego, La Jolla, CA, USA • Department of Biomedical Systems Informatics and Brain Korea 21 PLUS Project for Medical Science, Yonsei University College of Medicine, Seoul, South Korea • Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA, USA • Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA, USA • Department of Genetics, Stanford University, Stanford, CA, USA • Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA • Children's Medical Center Research Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA • Sarafan Chemistry, Engineering, and Medicine for Human Health (Sarafan ChEM-H), Stanford University, Stanford, CA, USA • Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA • Halıcıoğlu Data Science Institute, University of California at San Diego, La Jolla, CA, USA

Abstract

Extrachromosomal DNA is a common cause of oncogene amplification in cancer. The non-chromosomal inheritance of ecDNA enables tumors to rapidly evolve, contributing to treatment resistance and poor outcome for patients. The transcriptional context in which ecDNAs arise and progress, including chromosomally-driven transcription, is incompletely understood. We examined gene expression patterns of 870 tumors of varied histological types, to identify transcriptional correlates of ecDNA. Here we show that ecDNA containing tumors impact four major biological processes. Specifically, ecDNA containing tumors upregulate DNA damage and repair, cell cycle control, and mitotic processes, but downregulate global immune regulation pathways. Taken together, these results suggest profound alterations in gene regulation in ecDNA containing tumors, shedding light on molecular processes that give rise to their development and progression.

eLife assessment

This study of extrachromosomal DNA (ecDNA) identifies genes that distinguish ecDNA+ and ecDNA- tumors. The findings in the manuscript are **important** and the genomic analyses **convincing**. However, some of the data remain observational and the inferences would therefore be more robust with experimental validation. This manuscript could well be of relevance to biologists interested in cancer biology and gene regulation.

Introduction

Extrachromosomal DNA (ecDNA) are large, functional, circular double-stranded DNA molecules that are enriched for oncogenes, highly amplified, and frequently observed in a wide variety of cancer types[1,2]. ecDNAs lack centromeres and are asymmetrically segregated into daughter cells during cell division, driving intratumoral genetic heterogeneity, accelerated evolution, and rapid treatment resistance[3,4]. Further, recent studies demonstrate strong positive selection for ecDNA during tumor progression[5]. ecDNAs also exhibit highly accessible chromatin and altered cis- and trans-regulation, including cooperative intramolecular interactions[6], promoting elevated expression of oncogenic transcriptional programs[7–9], further contributing to poor outcome for patients[2].

The recent development of computational tools that enable detection of ecDNA from whole genome sequencing data, has facilitated analyses of well-curated, publicly available datasets, including The Cancer Genome Atlas (TCGA), thereby providing an important opportunity to identify transcriptional repertoires that are preferentially detected in bona fide, clinical ecDNA-containing tumors. To shed new light on the gene expression patterns that may enhance ecDNA development and progression, we examined global transcriptional analysis of ecDNA-containing tumors.

Results

A recent analysis utilized the tools AmpliconArchitect and AmpliconClassifier on 1,921 tumors from The Cancer Genome Atlas (TCGA) to suggest that ecDNA prevalence ranges from 0% to 59.6% across multiple tumor tissue subtypes[2]. Using AmpliconClassifier (AC), the analysis classified tumor samples into five subtypes: ecDNA(+), Breakage Fusion Bridge (BFB), complex non-cyclic, linear, and no-amplification. However, due to limitations imposed by short-read sequencing, AC may classify some ecDNA(+) structures as complex non-cyclic when breakpoints are missed. Secondly, BFB cycles can give rise to ecDNA formation, making discernment of the two modes of amplification difficult. To limit false-negative ecDNA classifications in the ecDNA(-) set, we treated samples with only a linear or no-amplification status as ecDNA(-), removing complex non-cyclic and BFB(+) samples from the analysis. In order to understand the transcriptional programs active in maintaining ecDNA, we selected 870 samples from 14 tumor types with at least three ecDNA(+) samples each, and compared the gene expression data of the resulting 234 ecDNA(+) and 636 ecDNA(-) samples (**Table S1**).

Machine learning identifies candidate genes for ecDNA maintenance

In lieu of identifying genes that are highly differentially expressed between ecDNA(+) and ecDNA(-) samples but driven by a small subset of cases (e.g. gene A in **Fig. S1a**), we sought to identify genes (e.g. gene B) whose expression level was predictive of ecDNA presence. We assumed that genes that were persistently over-expressed or under-expressed in ecDNA(+) samples relative to ecDNA(-) samples were more likely to be involved in ecDNA biogenesis or maintenance, or in mediating the cellular response to the presence of ecDNA.

To identify a minimal set of genes whose expression values were consistently predictive of ecDNA presence, we used Boruta,[10] an automated feature selection algorithm (**Fig. 1a** and Methods). Given the unequal representation of ecDNA(+) and ecDNA(-) samples within each of the 14 tumor types, we performed Boruta on 200 datasets, each consisting of a random selection of 80% of the

870 samples (**Fig. 1a**), and chose the criterion of a gene being labeled as a Boruta gene in at least 10 of the 200 trials to be selected for downstream analysis. The Boruta analysis identified a set of 408 genes with persistent differential expression, hereafter denoted as the Core gene set.

Extending the Core set with co-expressed genes

We note that the Core gene set is not a comprehensive list of discriminatory genes, using a toy example. Consider gene "B", a member of the core gene set, and another gene, "C", whose expression values across all samples are nearly identical to the expression values of core gene B. The Boruta analysis would not need to assign gene C to the core set in addition to gene B, because adding both genes incurs the same predictive power as adding one. However, either, or both genes may play an important functional role. To correct this, we ran pvclust[11] to cluster all gene expression values, and to identify stable clusters using multiscale bootstrap resampling (**Fig. 1b**; Methods). We used an approximately unbiased (AU) confidence value of 0.95 to select the most highly co-expressed gene clusters. An AU confidence value of 0.95 represents the rejection of the null hypothesis that a group of genes fail to form a stable cluster at a significance level of 0.05. Recomputing the number of Boruta trials that members of a cluster were selected in, we selected clusters that appeared in at least 10 of the 200 Boruta trials (Methods). This resulted in the selection of 354 recurring clusters (**Table S2**).

Notably, among the 354 clusters, only 2 clusters (with 14 total genes) did not contain any Core genes. As most genes do not have completely identical expression patterns, we would expect one gene to be consistently picked as a Boruta gene over another co-expressed gene. Consistent with this hypothesis, most (344/354) clusters contained only 1 or 2 Core genes (**Fig. 1c**). When selecting clusters that contained at least 1 Core and 1 co-expressed gene, 53 of 71 clusters contained 1 to 3 Core genes (**Fig. S1b**), confirming that a few genes per co-expressed cluster provide sufficient predictive value, but other co-expressed genes might still play an important functional role in maintaining ecDNA presence. This is true for clusters of various sizes, including the 2-member cluster #74 and the 21-member cluster #3. In cluster #74, *CSTF1* had similar expression values to the Core gene *RAE1*, which is a mitotic checkpoint regulator implicated in tumor progression[12] (**Fig. S1c; Table S3**). While not necessarily increasing the predictive value, *CSTF1* is also a proto-oncogene involved in aberrant alternative splicing events[13]. In cluster #3, 12 genes were highly co-expressed with 9 Core genes (**Fig. S1d; Table S3**), and were enriched in cell-cycle related biological processes (Methods). Importantly, the total number of genes per cluster was also small (**Fig. 1d**), with only 7 of 354 clusters carrying more than 10 genes. This suggests that the Core genes have specific roles that cannot be accomplished by multiple other genes.

Summarizing, the 354 clusters contained 643 genes, which included 408 Core genes and 235 additional genes (**Fig. 1b**). Together, we define these genes as the CorEx (Core+co-expressed) genes (**Table 1**). To address the concern that the selection of CorEx genes based on bulk RNA-seq expression data could be confounded by tumor purity, we utilized a composite tumor purity score (CPE)[14], and observed that the ecDNA(-) samples had slightly (but significantly) lower purity than ecDNA(+) samples (*p*-value 0.0036; **Fig. S2a**). This is consistent with reduced detection of ecDNA in less pure samples. However, lower sensitivity of ecDNA detection would reduce the strength of the signal but not result in false positives. Indeed, when we compared the significance of CorEx gene directionality in highly pure samples (tumor purity≥0.8; n=287) versus all samples (n=870), we found significant correlation (**Fig. S2b**), indicating robustness of the CorEx set. The remaining manuscript investigates the functional properties of these genes.

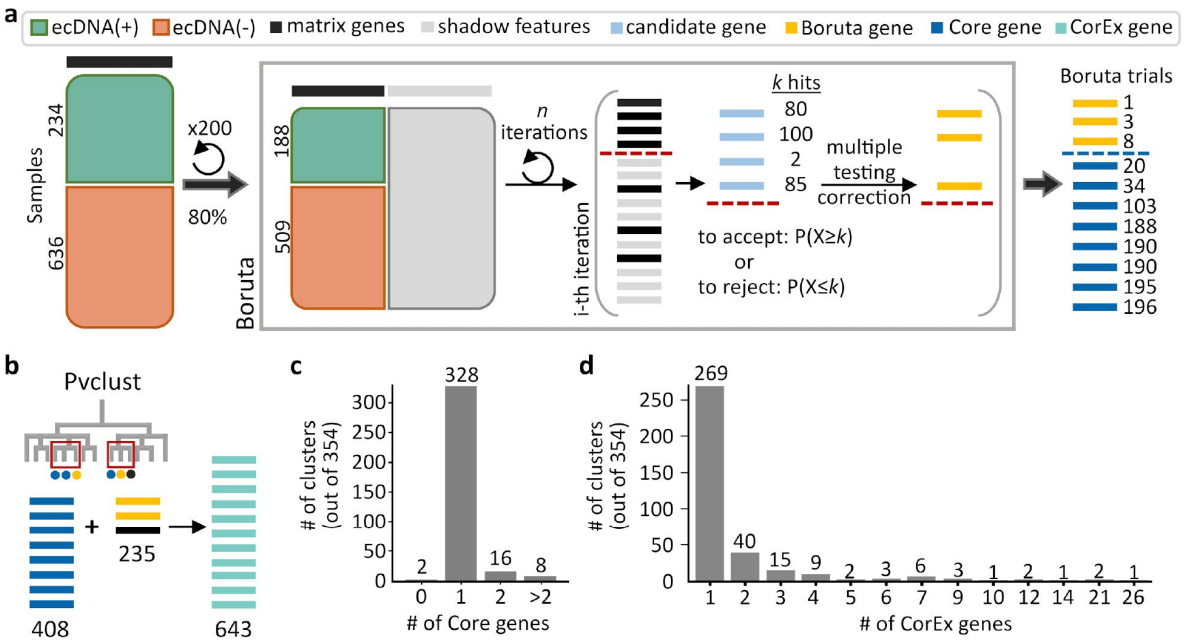


Figure 1.

Genes predictive of ecDNA status.

(a) The feature selection algorithm, Boruta, was applied to 200 datasets of randomly selected subsets consisting of 80% of all samples. Genes selected by Boruta in at least 10 of the 200 trials were identified as the Core set of genes (408) that were predictive of ecDNA presence. **(b)** Identification of highly co-expressed and stable gene clusters using pvclust expanded the Core set by an additional 235 genes to the final list of 643 CorEx genes. **(c)** Out of 354 clusters, the majority (344) of clusters contained 1 or 2 Core genes. **(d)** Most clusters were small, with only 7 clusters containing more than 10 genes.

CorEx genes are better predictors of ecDNA status compared to other gene sets

We validated the relevance of CorEx genes in ecDNA presence by running cross-validation experiments (**Fig. 2a**; Methods) to test the predictive power of CorEx gene expression in determining the ecDNA status of the sample. For comparisons, we used three other gene lists. The first list was a randomly chosen gene subset of identical size. For the second list, we performed a differential expression analysis using DESeq2[15] and picked the 643 most significantly differentially expressed genes in terms of the absolute value of their shrunken log-fold change estimate (LFC; Methods). Using the sign of the LFC value as the determinant for directionality, 240 of these genes were up-regulated, while 403 were down-regulated. Notably, only 86 of these Top-|LFC| genes overlapped with the CorEx gene set (**Table S5**; **Fig. S3d**). For the third list, we used a generalized linear model (GLM) to predict 3,012 genes whose expression levels were significantly associated with sample ecDNA status using a logit function after controlling for tumor subtype (Methods). Together, the 3 additional gene lists were denoted as random, Top-|LFC|, and GLM.

For cross-validation tests, we performed multiple random 80-20 splits of the samples to generate 200 training and test data-sets (**Fig. 2a**). For each training-test data-set, a Random Forest method was used to train the predictability of the 5 gene lists (Methods) on the training data, and the predictive performance was tested on the test data. Expectedly, none of the gene lists was a great predictor of ecDNA status of a sample. Nevertheless, the average of the area under the precision recall curve (AUPRC) was higher for CorEx and Core genes (0.48 and 0.5), relative to GLM and Top-|LFC| (mean AUPRC: 0.43 each; **Fig. 2b**, **S3b**). For precision values of at least 0.7, the CorEx genes had significantly higher recall than Top-|LFC| genes (Mann-Whitney U-test p-value 4.8e-21) or GLM genes (p-value 8.5e-20). In turn, the Top-|LFC| and GLM genes were more predictive than random (mean AUPRC: 0.36). Expectedly, the predictive performance did not change when switching between Core genes and CorEx genes, because each of the non-core gene in the CorEx list had an expression pattern similar to at least one Core gene (**Fig. 2b**, **S3a**).

To test the persistence of CorEx genes across tumor types, we re-computed Cliff's delta values[16,17] for each of the 11 TCGA tumor types that had at least 10 ecDNA(+) and at least 10 ecDNA(-) samples. The directionality of gene expression patterns was significantly similar to TCGA in each tissue type, with one exception (**Fig. 2c**; **Table S4**). The sole exception was the tumor type of Sarcoma (SARC). It is notable that the TCGA-SARC samples included many liposarcomas. In addition to containing ecDNA, liposarcoma samples are known to have extensively rearranged structures indicative of chromothripsis and neo-chromosome formation[18]. For other tissue types, the *p*-values against a null hypothesis of no match to the pan-cancer prediction ranged from 4.2e-12 to 3.5e-85 (Fisher's exact test) for the significant associations (Methods). The results were similar if we tested using only Core genes (**Fig. S3c**). Summarizing, the 643 CorEx genes are differentially expressed across a multitude of tumor types, and have consistently higher or lower expression in ecDNA(+) samples relative to ecDNA(-) samples. These results are consistent with a pan-cancer role of CorEx genes in ecDNA biogenesis and maintenance.

The Top-|LFC| genes were also different from the CorEx genes by other metrics. Not surprisingly, the log-fold change (LFC) values of the top-|LFC| genes were higher than the LFC values of the CorEx genes (**Fig. 2d**, MWU *p*-value 1.83e-158). However, much of the LFC change was due to the very low expression of the top-|LFC| genes in either ecDNA(+), or ecDNA(-) samples. In fact, the CorEx genes had higher expression in both ecDNA(+) and ecDNA(-) samples compared to the Differentially Expressed (DE) genes (**Fig. 2e**, MWU *p*-value < 2e-308). While the absolute log fold-change in expression of CorEx genes between ecDNA(+) and ecDNA(-) samples was not that high (median: 0.30, mean: 0.41), it was persistent across all samples (variance: 0.14, standard deviation: 0.37).

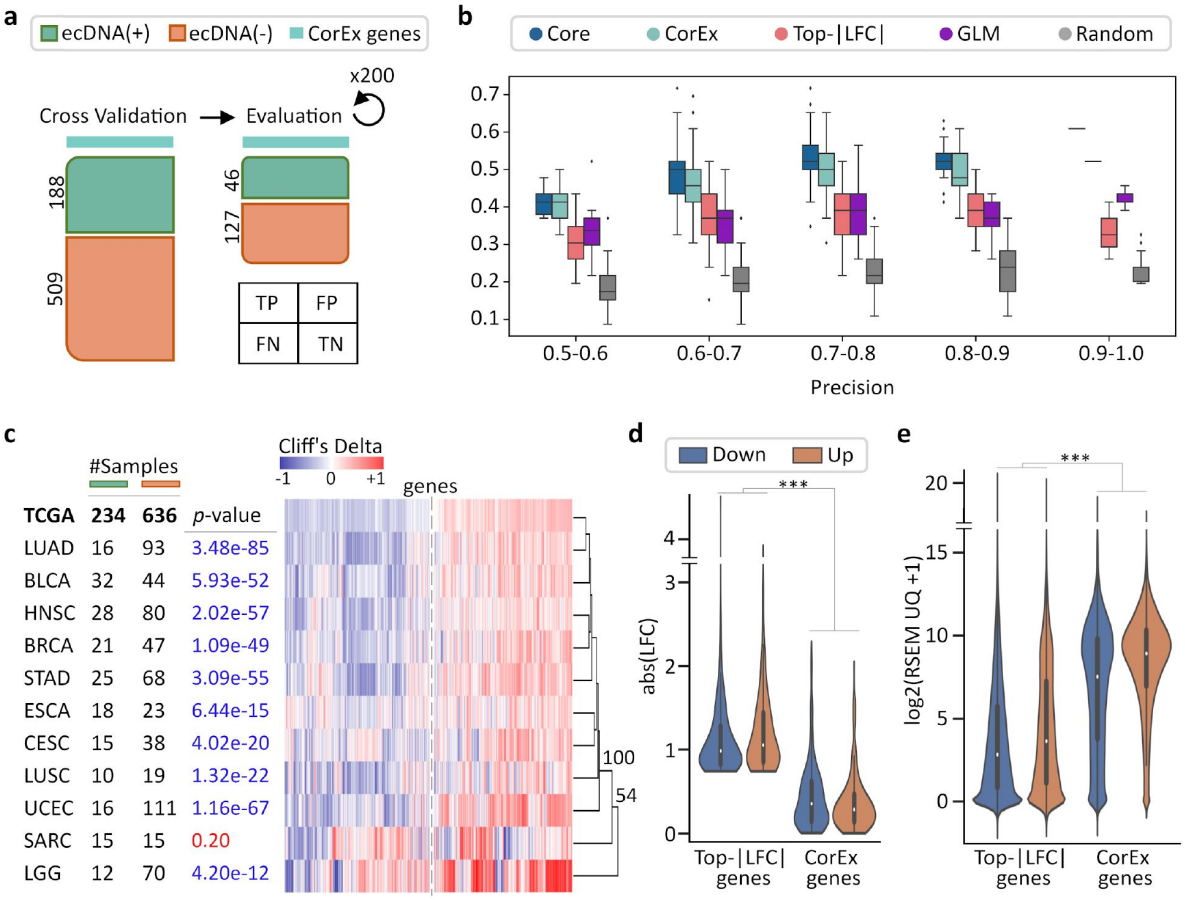


Figure 2.

Validation of CorEx genes.

(a) Cross-validation experiments validating the predictive value of CorEx genes. Precision denotes the fraction of predicted samples that were truly ecDNA(+). Recall refers to the fraction of ecDNA(+) samples that were predicted correctly. **(b)** For precision windows of width 0.1 and a value of at least 0.5, recall values were plotted as boxplots. The interquartile ranges for CorEx and Core genes overlap, suggesting similar predictive power. (Continued on the following page.). CorEx genes have higher predictive rates compared to the top 643 differentially expressed genes based on logarithmic fold changes from a DESeq2 analysis (Top-|LFC| genes), 3,012 significant genes selected from a generalized linear model (GLM), and 643 randomly selected genes. **(c)** CorEx genes were consistently up- or down-regulated in ecDNA(+) samples across tumor types, with the exception of SARC. AU *p*-values from multiscale bootstrap resampling are shown at the dendrogram branches. **(d)** Of the 643 Top-|LFC| genes, 240 were up-regulated while 403 were down-regulated in ecDNA(+) samples. Of the CorEx genes, 325 were up-regulated while 318 were down-regulated. The absolute LFC values of the Top-|LFC| gene set was significantly greater than that of the CorEx genes (*p*-value 1.83e-158). **(e)** The normalized gene expression values of the CorEx genes were significantly higher than that of the Top-|LFC| gene set (*p*-value < 2e-308). ***p-value < 0.001.

For example, the genes *ITLN1* and *PNMT* had the second and eighth-highest absolute LFC values of 3.92 and 2.89 in the top-|LFC| list. However, their normalized expression values in most ecDNA(+) samples were low. *ITLN1* had a normalized RSEM expression value ≤ *8* (21st percentile) in 210/234 ecDNA(+) samples. Similarly, the normalized RSEM expression value of *PNMT* in 223/234 ecDNA(+) samples was less than 8.5 (rank percentile: 41.1%). For *PNMT*, the differential expression was mediated by 11 ecDNA(+) samples having an expression value ≥ *11*, and 5 of the 11 samples contained *PNMT* on an ecDNA amplicon (**Fig. S3e**). Similarly, 3 samples with high RSEM contained *ITLN1* on an amplicon (**Fig. S3f**), partly accounting for the high |LFC| value. In contrast, the CorEx gene, *RAE1*, had a high normalized expression value in both ecDNA(+) and ecDNA(-) samples (average 9.72, rank percentile 74.3%), with a small but persistent LFC value of 0.33.

The results confirm our intuition that differential expression can arise due to multiple reasons, including low expression of the gene in a majority of samples, or the copy number amplification of a gene in a few samples. In contrast, the CorEx genes were selected based on persistent over- or under-expression in ecDNA(+) samples.

CorEx genes primarily up-regulate three biological processes: Cell Cycle, Cell division, and DNA Damage Response

To identify enriched biological processes specific to either up-regulated or down-regulated genes in ecDNA(+) samples, we combined two metrics of effect size, Cliff's delta[16,17], and log fold change[15] to determine the directionality of CorEx genes (**Table S6**; Methods). The two effect size metrics were mostly in agreement in terms of directionality. Of the 7,288 genes that passed the negligible effect size thresholds in both metrics, only 14 were not in concordance. This more stringent approach, in comparison to a simple directionality based on the sign of a single effect size value, was applied given that enrichment analysis on gene sets is dependent on not only the number of up- or down-regulated genes but also the degree of overlap with genes under a specific biological process term (Methods). Using this approach, among the 643 CorEx genes, 262 genes were found to be up-regulated in ecDNA(+), while 271 were found to be down-regulated (**Table 1**; Methods). 110 genes did not make the effect size cut-off. The numbers were similar for the 408 Core genes, with 190 up-regulated, 196 down-regulated, and 22 genes not making the cut-off.

We performed enrichment analysis on gene sets to identify the Gene Ontology (GO) biological processes that are enriched in CorEx genes (Methods). Briefly, we applied a one-sided Fisher's exact test using gene sets from MSigDB[19–21], using a false discovery rate of 5% (Benjamini-Hochberg procedure). The UP-regulated genes enriched 187 Biological processes (**Table S7**). Note that the GO-biological process (BP) terms are not independent, because of their hierarchical organization, and sharing of genes across different GO terms. Therefore, we used an approach similar to that used in DAVID[22] to cluster the biological processes enriched by the UP-regulated genes into 11 broad categories (**Table S8**; **Fig S4**; Methods). The 11 categories were assigned a name using manual inspection of the constituent GO terms, or called "Other." The 11 categories (including "Other") are shown in a waterfall plot to explain the contribution of each gene to a category (**Fig. 3a**).

The 10 categories included expected participation of biological processes involved in (a) cell-cycle regulation (Mitotic/Meiotic Cell Cycle, G1/S, G2/M) (b) cell-division (Spindle Organization, Cell Division, Chromosome Condensation, Chromosome Segregation), (c) DNA Damage response (DNA Repair), and (d) the HOX Gene cluster. Indeed, one of the largest clusters, cluster #3, containing 9 Core genes and 12 co-expressed genes, was enriched in GO-BP terms related to the cell cycle (**Table S9**). Notably, the enriched categories also included a role for the *HOX* genes with 17 members of the *HOX* family up-regulated in ecDNA(+) cancers (**Table S8**). Many recent reports have associated *HOX* genes with cancer, including an association with their phenotypic "hallmarks"[23]. Genes

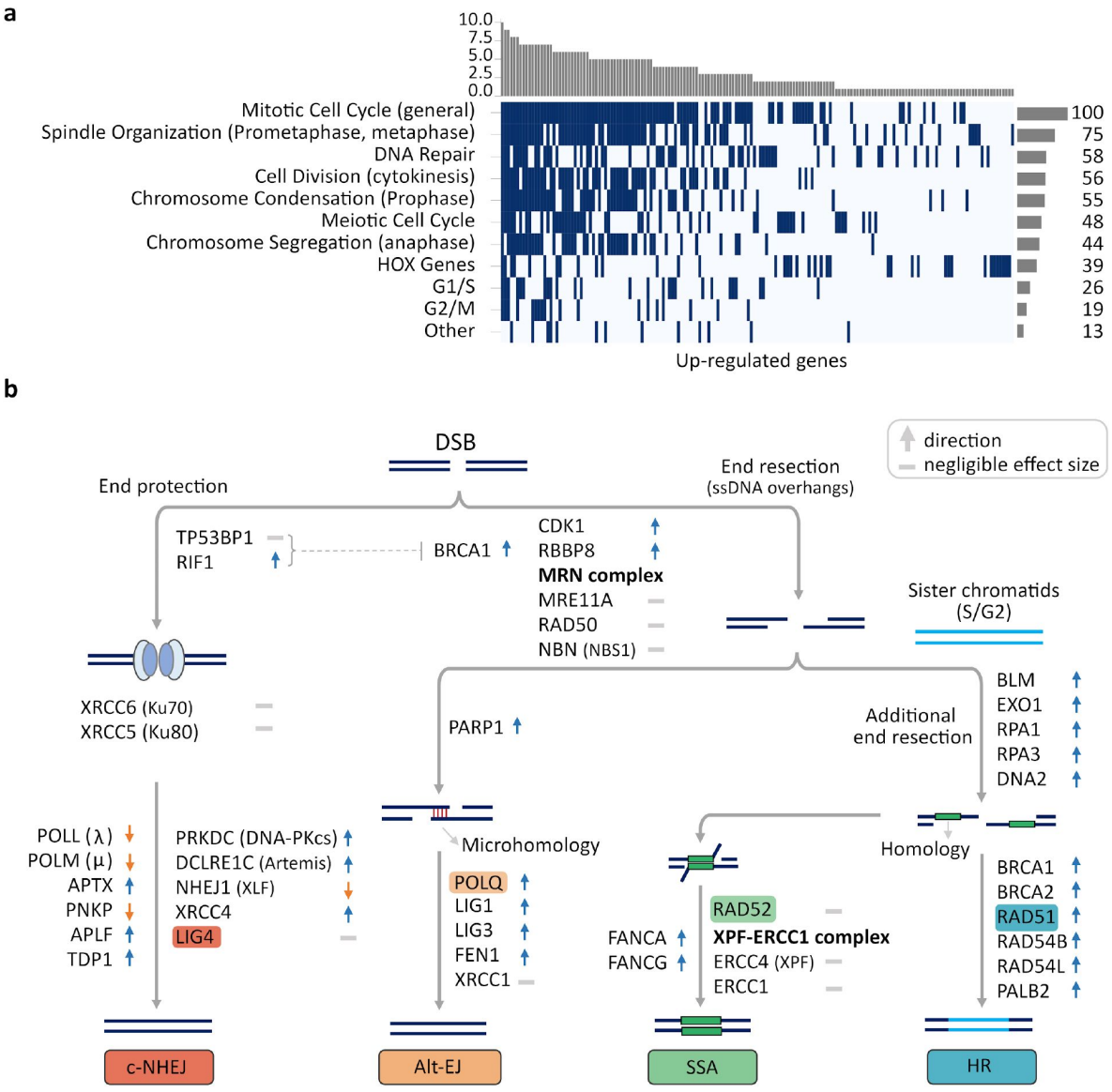


Figure 3.

Up-regulated CorEx genes.

(a) GO biological processes enriched in up-regulated genes were clustered into 11 broad categories. The horizontal barplot represents the number of GO biological processes belonging to each of the 11 broad categories, while the vertical barplot represents the number of broad categories that a specific GO biological process belongs to. **(b)** Genes up- or down-regulated in processes involved in major double-strand break (DSB) damage repair pathways. Many critical genes in the c-NHEJ pathway were down-regulated in ecDNA(+) samples relative to ecDNA(-) samples.

involved in angiogenesis (*HOXA2*, *HOXC5*), genome instability (*HOXC5*, *HOXC11*), deregulating cellular energetics (*HOXA4*, *HOXC5*), and metastasis (*HOXA2*) were all up-regulated in ecDNA(+) cancers.

While Meiotic cell-cycle was also enriched, only 7 genes were allocated specifically to the group of enriched terms: *SEPP1*, *SNRPA1*, *TMEM203*, *RNF114*, *TAF4*, *TNFAIP6*, and *PTX3* (**Table S10**). Though these genes have roles in the meiotic cell cycle, they have also been implicated in cancer and other inflammatory diseases. The spliceosomal protein SNRPA1 is a pro-metastatic splicing enhancer[24]. TNFAIP6, together with PTX3, activates the Wnt/β-catenin pathway to promote gastric carcinoma cell invasion[25]. TMEM203 is a STING-centered signaling regulator implicated in inflammatory diseases[26]. RNF114 is a zinc-binding protein whose over-expression is an indicator of epithelial inflammation and implicated in various tumors[27]. TAF4, a transcription initiation factor, when overexpressed, is implicated in ovarian cancer by playing a role in dedifferentiation that promotes metastasis and chemoresistance[28]. Finally, in looking at the genes in the "Other" category, *SHCBP1* was the only gene unique to it. A member of the neural precursor cell proliferation process, SHCBP1 is reported to promote tumor cell signaling and proliferation[29].

Taken together, the 11 biological process categories explain 165 of the 262 up-regulated CorEx genes, and suggest that Mitotic Cell-Division, Cell cycle regulation, and DNA Damage response are the three broad categories of biological processes up-regulated to ensure ecDNA presence, along with an up-regulation of genes in the *HOX* cluster.

CorEx genes upregulate specific Double-strand break repair pathways

The 16 enriched DNA damage response GO terms contained terms "double-strand break repair" and "recombination", but none contained the terms "single-strand", "nucleotide-excision", or "mismatch-repair" (**Table S8**), suggesting that the CorEx genes are largely composed of genes involved in multiple double-strand break (DSB) repair pathways, which include classical non-homologous end-joining (c-NHEJ), Alternative end-joining (Alt-EJ), single-strand annealing (SSA), or homology directed repair (HDR)[30].

The choice of these varied DSB repair mechanisms for ecDNA presence is not well understood. We compiled and hand-curated a list of 129 genes involved in DSB repair and marked them for their role in one or more of these 4 pathways (**Table S11**). Of these genes, a high number (67) were up-regulated in ecDNA(+), while a smaller number (15) were down-regulated, relative to ecDNA(-) samples. This breakdown of 129 DDR genes contrasts with an analysis using all genes where a near identical number of genes (5,256, and 5,251) were up- and down-regulated in ecDNA(+) samples, confirming that DDR genes are significantly up-regulated relative to all differentially expressed genes (p-value < 0.0001; Fisher's exact test). When broken down to the roles of genes in individual DSB repair pathways, we found that Alt-EJ with 11 up-regulated and 1 down-regulated genes (p-value 0.0063), SSA (11 up, 1 down (p-value 0.0063)), and HR (46 up, 8 down; p-value <0.00001) were all up-regulated. However, classical NHEJ (14 up, 7 down; p-value: 0.19) was not significantly up-regulated in ecDNA(+) samples relative to ecDNA(-) samples (Methods, **Table S12, Table S13**).

The expression of key genes in these pathways raises the possibility of an increased role of non-classical-NHEJ processes in ecDNA development or progression, relative to c-NHEJ (**Fig. 3b**; **Table S11**). A number of genes involved in c-NHEJ were downregulated in ecDNA-containing tumors relative to non-ecDNA tumors. These included *XLF*/*NHEJ1* (MWU p-value 2.05e-03), which is a key member of the ligase complex required for c-NHEJ; *LIG4*, another member of the ligase complex (MWU p-value 0.03), *PNKP*, which generates 5′-phosphate/3′-hydroxyl DNA termini required for ligation (MWU p-value 3.90e-06); and also, DNA polymerases λ (*POLL*; MWU p-value 1.09e-21) and μ (*POLM*; MWU p-value 0.01), which promotes the ligation of terminally compatible

overhangs requiring fill-in synthesis and promotes the ligation of incompatible 3′overhangs[31] in a template independent manner, respectively. This does not imply a defect in these repair processes, but rather, potentially additional or preferential utilization of alternative DSB repair pathways in ecDNA-containing tumors.

TP53BP1 is key to blocking resection and promoting the c-NHEJ pathway choice, but is displaced by BRCA1 and the MRN complex to initiate resection in the broken strands[32,33]. *BRCA1* was significantly up-regulated in ecDNA(+) samples (**Fig. 3b**), while *TP53BP-1* was significantly down-regulated (MWU *p*-value 0.016), although with negligible effect size (**Table S11)**. Supporting the role of alternative pathway choice for DDR, key genes in the Alt-EJ pathway, including *PARP-1*, DNA polymerase θ (*POLQ*), *LIG1*, *LIG3*, *FEN1* were all significantly up-regulated in ecDNA(+) samples. Homology directed repair is the preferred pathway when a sister chromatid is available to act as a template. HDR is initiated by additional and extensive resection. The genes *BLM*, *EXO1*, *RPA1*, *RPA3* which promote additional resection, as well as *BRCA1*, *BRCA2*, *RAD51*, and others that support HDR were all found to be significantly up-regulated. We can conclude that the specific pathway choice for DSB repair in ecDNA(+) samples is dominated by alt-EJ and homology directed repair pathways, while c-NHEJ is not a preferred choice.

CorEx genes primarily down-regulate immune system processes

Using methodology similar to the analysis of the up-regulated genes, the down-regulated genes enriched 73 GO terms (**Table S14**), and could be clustered into seven broad categories, including "Other" (**Fig. 4a**; **Table S15**; **Fig. S5**). Surprisingly, all categories were immunomodulatory. The most enriched broad category contained 75 CorEx genes relating to the Lymphocyte activation pathway. It included genes enriching "T-cell activation" (28 CorEx genes; *p*-value 2.58e-05), and "Positive regulation of cell-cell adhesion" (16 CorEx genes; *p*-value 4.49e-03). Other down-regulated pathways included Cytokine activation, especially for genes in the IL-12 pathway (6 CorEx genes, *p*-value 7.17e-03), TNF super-family (12 CorEx genes, *p*-value 7.24e-03), and Inflammation, including, for example, down-regulation of Toll-like receptor 2 signaling (4 CorEx genes, *p*-value 4.49e-03). Finally, the broad category of Leukocyte chemotaxis was also enriched among the down-regulated genes. The chemotaxis genes include many chemokines and their receptors involved in trafficking of T cells to the site of the tumor. The remaining down-regulated genes included four fucosyltransferases, and the category marked "Other." *FUT2* silencing is associated with reduced adhesion and increased metastatic potential[34]. Notably, the category marked "Other" was dominated by genes in NF-κB pathway regulation (14 CorEx genes, *p*-value 2.28e-02).

NF-κB signaling represents a prototypical, proinflammatory pathway[35] with multiple roles, including apoptosis. Specifically, 8 of the 14 down-regulated genes involved caspase activation (**Table S16**), representing the pro-apoptotic arm of NF-κB signaling. A parallel pathway for sensing endogenous ligands secreted in cell death and cancer is mediated by Toll-like receptor (TLR) proteins[36]. Remarkably, all ten TLRs were significantly down-regulated in ecDNA(+) tumors. They included TLRs expressed on the cell membrane that bind lipids and proteins as well as TLRs expressed on endosomal membranes that bind DNA. The CorEx down-regulated genes also included many involved in TLR signaling, such as *TLR3*, *CYBA*, *LYN*, and *TIRAP*.

Four of the seven broad categories mapped to facets of the cancer immune cycle[37] (**Fig. 4b**). We tested if CorEx genes in these categories were more likely to be down-regulated rather than up-regulated, when compared to the non-CorEx differentially expressed genes. The Inflammation category, which mapped to the "Cancer antigen presentation" facet, showed 18 up-regulated and 76 down-regulated CorEx genes (*p*-value 0.005, Fisher exact test, **Table S15**). Similarly, the CorEx genes related to the "Trafficking of T cells" facet ("Leukocyte migration and chemotaxis" category, 13 up, 49 down-regulated; *p*-value 0.03) and "Infiltration and recognition of tumor cells by cytotoxic T cells" facet ("Lymphocyte activation" category, 23 up, 75 down-regulated; *p*-value

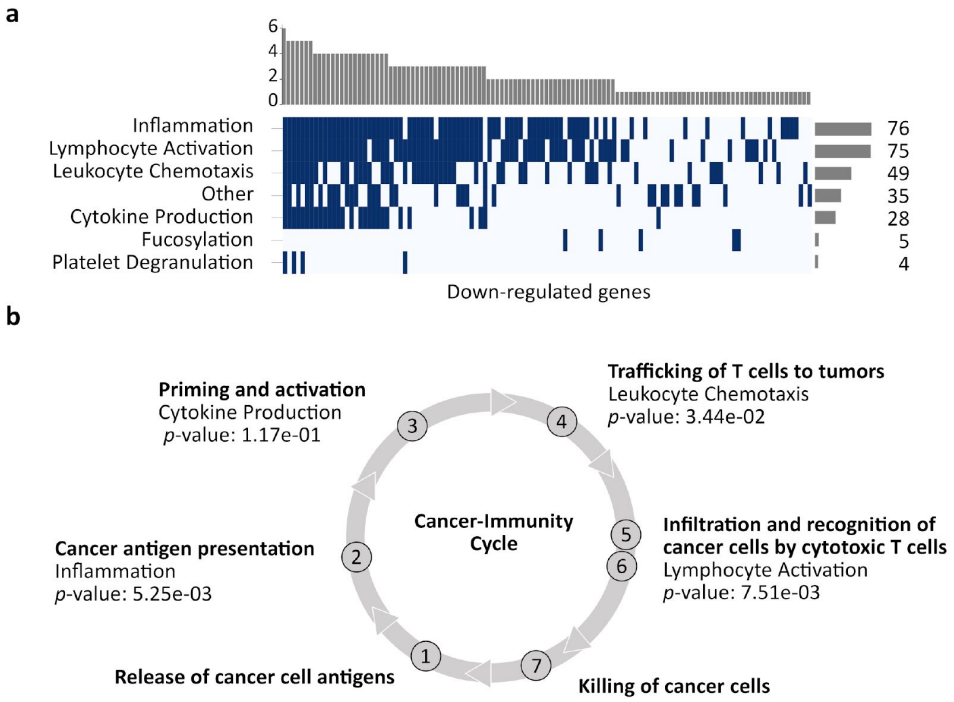


Figure 4.

Down-regulated CorEx genes.

(a) GO biological processes enriched in down-regulated genes were clustered into 7 broad categories. The horizontal barplot represents the number of GO biological processes belonging to each of the 7 broad categories, while the vertical barplot represents the number of broad categories that a specific GO biological process belongs to. **(b)** Four of these categories map to steps in the cancer-immunity cycle. CorEx genes in three of the four categories were significantly down-regulated compared to all genes (Fisher's exact test).

0.0075) were also significantly down-regulated. However, down-regulation in the "Priming and activation" facet ("Cytokine production" category, 6 up, 28 down-regulated) was not significant at the 5% level.

As the RNA data were bulk-sequenced, transcripts were sampled from tumor cells and cells from the tumor microenvironment. Thorsson *et al*.[38] mined immune cell expression signatures to identify six immune subtypes: wound healing (C1), IFN-γ dominant (C2), inflammatory (C3), lymphocyte depleted (C4), immunologically quiet (C5), and TGF-β dominant (C6). A recent study analyzing the tumor microenvironment (TME) of ecDNA(+) vs. ecDNA(-) samples in seven tumor subtypes revealed an association of ecDNA presence with immune evasion[39]. Our results (**Fig. S6**), which used an updated version of the classification method for these ecDNA(+) samples, were broadly consistent with those from the Wu *et al*[39]. study. Our results suggested an increase in C1 and C2 subtypes and a depletion of C3 and C6 between ecDNA(+) and ecDNA(-) categories (*p*-value 3.96e-03, Chi-squared test). Notably, the C3 (inflammatory) subtype is associated with lower levels of somatic copy number alterations, and C6 with high lymphocyte infiltration, while C1 is associated with elevated levels of angiogenic genes. These are consistent with our findings of increased somatic copy numbers, increased expression of angiogenic genes on ecDNA(+) samples, and reduced lymphocyte infiltration.

ecDNA(+) samples carry a higher mutational burden relative to ecDNA(-) samples

In order to understand if the change in transcriptional program was driven by mutations to the genes, we checked if ecDNA(+) samples have differential levels of mutation relative to ecDNA(-). Intriguingly, we found that the total mutation burden was significantly higher in ecDNA(+) samples relative to ecDNA(-) samples (**Fig. 5a**). The result was significant also when mutations were limited to deleterious substitutions as measured by SIFT or PolyPhen2, and high-impact insertions and deletions (**Fig. S7a**). However, when controlling for cancer type, only glioblastoma (GBM; lower mutations in ecDNA(+)), low-grade gliomas (LGG; higher mutations in ecDNA(+)), and uterine corpus endometrial carcinoma (UCEC; lower mutations in ecDNA(+)) continued to show differential total mutational burden (**Fig. S7b**). Next, we tested if specific genes were differentially mutated between the two classes (**Fig. 5b**). For deleterious/high-impact mutations, *TP53* was the only gene whose mutational patterns were significantly higher in ecDNA(+) compared to ecDNA(-) (OR 2.67, Bonferroni adjusted *p*-value 4.22e-07). BRAF mutations, however, were more common in ecDNA(-) samples and were significant to an adjusted *p*-value < 0.1 (OR 0.27). The excess of *TP53* mutations in ecDNA(+) samples provides additional support to the hypothesis that mutations in DNA damage response or cell cycle checkpoints are important for ecDNA presence. Other genes that are differentially mutated with nominal significance (unadjusted *p*-value < 0.005) are shown in **Table S17**.

We also tested if a collection of gene mutations could predict ecDNA status using XGBoost[40], which uses an adaptive boosting of "weak classifiers" to predict class. Here, each mutated gene was treated as a weak classifier of ecDNA status. However, the two classes could not be separated with high accuracy (**Fig. S7c**). An unsupervised principal component analysis did not separate the two classes either. Only the first principal component explained a significant proportion (14%) of the total variance (**Fig. S7d**) and did not separate the bulk of the samples. Finally, we recapitulated earlier findings that ecDNA(+) samples enrich for APOBEC activity through the presence of the mutation signatures SBS2 and SBS13[41,42] (**Fig. S8**). The enrichment in *TP53* mutations was also consistent with previous findings[5]. On balance, however, collections of gene mutations did not distinguish ecDNA(+) samples from ecDNA(-) samples, at least at a pan-cancer level, in contrast to the gene expression data.

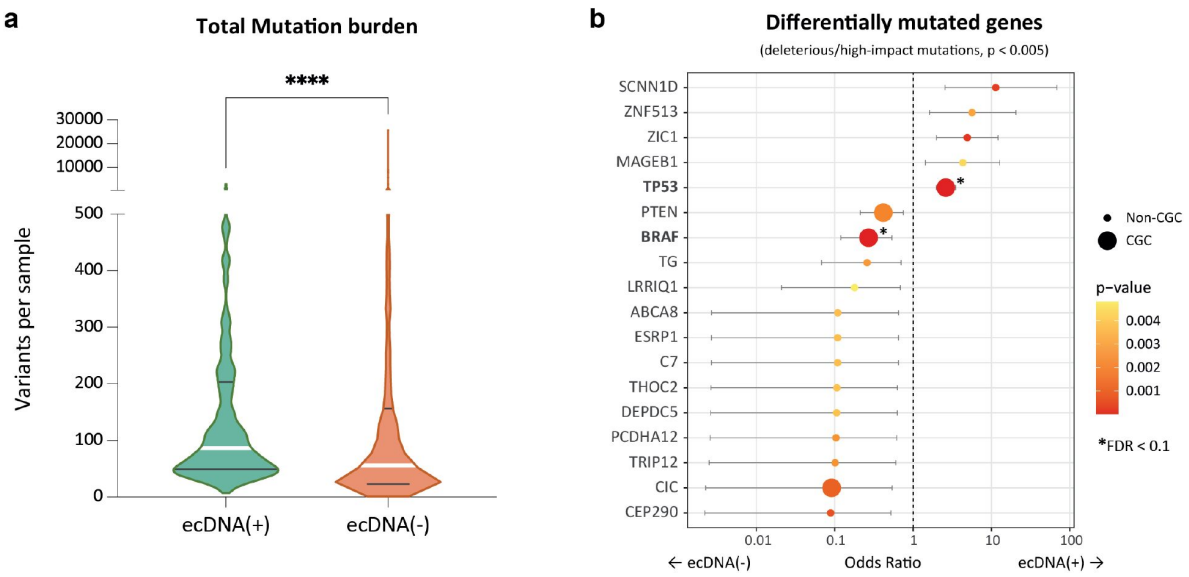


Figure 5.

Mutational characteristics of ecDNA-containing tumors.

(a) Total mutation burden of ecDNA(+) and ecDNA(-) samples. ecDNA(+) samples have significantly higher mutation burden than the ecDNA(-) samples (p-value < 0.0001, Mann Whitney test). **(b)** Odds ratios of differentially mutated genes in ecDNA(+) and ecDNA(-) (p-value < 0.005). The size of the dot indicates whether the corresponding gene belongs to the Cancer Gene Census (CGC) or not (Non-CGC). Only *TP53* and *BRAF* showed significance at the level of FDR < 0.1 (Benjamini-Hochberg).

Persistently occurring genes in ecDNA(+) samples represent potential vulnerabilities

Any CorEx gene is either a Core gene that was selected as a feature in at least 5% of 200 Boruta trials, or be highly co-expressed with a Core gene. Because the selection criterion of 5% is arbitrary, we also tested robustness with 8 other cut-offs ranging from 5-of-200 to 200-of-200 Boruta trials. The number of CorEx genes expectedly decreases with more stringent cut-offs. However, of the 187 GO terms that were enriched by 262 CorEx UP-genes using 10 of 200 Boruta trials as the selection criteria, 93 terms (49.7%) were enriched for each cut-off (**Fig. S9**), and 155 terms (82.9%) were enriched in at least 5 of the 8 cut-off criteria. Given that our subsequent analyses utilized the hierarchy of GO terms and identified 4 GO-categories enriched by UP-regulated genes, the conclusions would hold regardless of the specific cut-off.

To rank CorEx genes by importance, we computed harmonic mean rank values based on three categories: a) the average GINI importance statistic from the trained random forest models; b) the number of Boruta trials that a gene was selected in; and c) the number of Boruta trials (out of 200) that a gene was selected in when counting by cluster (Methods). 65 genes that were up-regulated (47 genes) or down-regulated (18 genes) had a harmonic rank lower than 3 (**Table 1)**. The next highest ranked gene had harmonic rank exceeding 17. These 65 genes represent the most persistent differentially expressed CorEx genes, and appeared as Core (or clustered gene) in all 200 Boruta trials. Notably, of the 24 genes most frequently expressed on ecDNA,[2] only EGFR and CDK4 were included in the list of 65 genes, suggesting that the most persistent CorEx genes do not themselves appear frequently on ecDNA.

Expectedly, the high-ranked up-regulated genes impacted cell division (16 genes), cell cycle regulation (10 genes), DNA damage response (16 genes). Only 12 of the 47 genes were not included in the gene sets of any enriched GO term. Many of these genes were from small CorEx clusters with less than 3 members, but we also found 6 genes from the HOX gene cluster (cluster #17), and another cluster of 21 genes (cluster #3). Members of cluster #3 appeared in all 200 Boruta trials; however, there were three genes all involved in cell-division (*TPX2*, *KIF2C*, and *AURKA*), each of which appeared in at least 180 Boruta trials. High expression among these three genes is associated with poor prognosis[43], and due to the highly persistent nature of their differential expression across ecDNA(+) samples, they represent a possible widespread vulnerability for ecDNA(+) samples.

Intriguingly, 14 of the 18 down-regulated genes with low harmonic rank came from a single cluster (#2; **Table S2**), and 13 of the 18 genes did not specifically enrich any specific BP ontology. Six of the down-regulated genes appeared in 180 or more Boruta trials (*CHMP7*, *XPO7*, *INTS9*, *TACR1*, *KIAA1967*, and *PCM1*). Some of these genes (*CHMP7*, *XPO7*, *KIAA1967*) are reported to be tumor suppressor genes[44–46]. However, the exact functional role of down-regulating these genes in ecDNA(+) samples remains to be elucidated.

Discussion

ecDNA is increasingly recognized as a major cause of oncogene amplification, intratumoral genetic heterogeneity, accelerated evolution, and treatment resistance, but many of the underlying processes involved in its formation, function, and progression are not fully understood. The ability to conduct multi-omic studies of well-curated, bona fide clinical tumor samples, such as the TCGA, presents an opportunity to learn about differentially regulated gene expression programs that may be involved in ecDNA biogenesis or maintenance, and in worse outcomes for patients[7–9]. Using a relatively intuitive set of principles, we have developed a machine learning approach that

identifies differentially expressed, co-regulated genes in ecDNA-containing tumors, highlighting four main biological processes: non-c-NHEJ DSB repair, cell cycle, proliferation control, and immune regulation.

The GO analysis revealed three core biological processes that were up-regulated and only the immune system processes as being down-regulated. These observations strengthen the case for targeting proteins involved in mitotic cell-division[47], cell-cycle regulation, and DNA damage response in ecDNA(+) cancers, but also reveal roles for the *HOX* cluster of genes. Also, in this paper, we did not extensively study the role of ncRNA in prediction of ecDNA status. We do note that *HOTAIR*, encoded in the *HOXC* locus, is independently associated with metastasis and poor outcomes[48]. Further experiments are needed to provide a mechanistic basis for the role of *HOX* cluster genes in maintaining ecDNA presence, as also for involvement of ncRNA.

The DNA damage genes are broadly up-regulated in ecDNA(+) samples, especially in double-strand break repair. Within this broad category of mechanisms, our analysis suggests that alternative DSB repair pathways such as Alt-EJ are preferred relative to classical NHEJ. This is consistent with previous observations of small microhomologies at breakpoint junctions[2,49], and has important implications in therapeutic selection that will need to be validated in future experimental studies. We note, however, the microhomology analyses typically study breakpoint junctions, and might ignore double-strand breaks in non-junctional sequences which could be observed, for example at replication-transcription junctions.

The down-regulated genes were primarily immunomodulatory in nature, in addition to a few persistently down-regulated tumor suppressor genes. Lowered expression of immunomodulatory genes in ecDNA(+) samples has been previously reported[2,39], but not mechanistically explained. Remarkably, the down-regulated immunomodulatory genes encompassed most aspects of the cancer immune cycle, suggesting impaired recognition of tumor DNA and proteins as foreign in ecDNA(+) tumors. Sensing of foreign DNA, including tumor DNA, is often mediated by the cGAS/STING pathway[50,51]. Intriguingly, cGAS was significantly up-regulated in ecDNA(+) samples, while STING was significantly down-regulated, suggesting a role for STING agonists in intervention. Finally, in addition to the down-regulation of genes in the toll-like receptor family, we observed a down-regulation of genes involved in regulating TLR signaling pathways that were part of the CorEx list. Understanding the mechanisms of broad down-regulation of TLRs could provide insight into vulnerabilities of ecDNA(+) tumors.

Mutation data alone does not provide as clear a picture of the genes involved in ecDNA status prediction. We did observe that the total mutation burden (TMB) was higher in ecDNA(+) samples. However, that relationship is much less clear after controlling for cancer type. High TMB has been positively correlated with sensitivity to immunotherapy[52], and better patient outcomes; however, the gene expression patterns suggest that immunomodulatory genes are down-regulated in ecDNA(+) samples, and patients with ecDNA(+) tumors have worse outcomes[2]. Notably, other results have suggested that the correlation between TMB and response to immunotherapy is not uniform, and it can vary across different tumor subtypes[53]. Specifically, our data is consistent with previous results which showed that Gliomas with high TMB have worse response to immunotherapy relative to gliomas with low TMB[53]. In general, no collection of gene mutations was predictive of ecDNA status, although mutations in *TP53* were more likely in ecDNA(+) samples, and perhaps are an important driver for ecDNA formation[5].

These results suggest that cancer cells that contain ecDNA have profound alterations in their global transcriptional patterns. Importantly, these transcriptional differences do not arise solely from genes on the ecDNAs themselves, but rather suggest that fundamental global processes involved in DSB repair, cell cycle control, and immune regulation contribute to ecDNA formation and pathogenesis.

Methods

TCGA sample ecDNA status classification

Amplicon Classifier (version 0.4.9, *https://github.com/jluebeck/AmpliconClassifier*) classified amplicons detected in 1,921 TCGA samples into five sub-types: ecDNA, BFB, complex non-cyclic, linear, and no-amplification. When classifying a sample with multiple amplicons, the order of preference is as follows: ecDNA, BFB, complex non-cyclic, linear, and no-amplification. Given the challenges of detecting ecDNA from short read data, and to avoid possible false-negative ecDNA classifications, samples with a BFB or complex non-cyclic status which were not called ecDNA(+), were removed from the analysis. We treated samples with the linear amplification and no-amplification classifications as ecDNA(-). Of the 1,921 samples, 1,535 samples classified as ecDNA(+) and ecDNA(-) had RNA-seq data, including 1,406 primary solid tumor samples, 95 tumor metastasis samples, and 34 primary blood derived cancer – peripheral blood samples. Removing metastases results in a total of 1,440 samples, including 243 ecDNA(+) and 1,197 ecDNA(-) samples. While the set of 1,440 samples represented 24 tumor types, ten of these tumor types had insufficient numbers of ecDNA(+) samples, including four tumor types with no ecDNA(+) samples. To prevent the 561 ecDNA(-) samples representing these tumors from skewing the analysis, we removed 570 samples representing tumor types with less than three ecDNA(+) samples. This resulted in a total of 870 samples representing 14 tumor types, of which 234 were classified as ecDNA(+) and 636 were classified as ecDNA(-).

Gene expression datasets

Gene expression data for 32 studies part of the TCGA Pan-cancer Atlas was downloaded from cBioPortal (01.05.2021) (*https://www.cbioportal.org/*). The cBioPortal "data_RNA_Seq_v2_expression_median.txt" data is sourced from the file "EB++AdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv" (synapse id: syn4976363). Briefly, the matrices contain batch corrected values of the upper-quartile (UQ) normalized RSEM estimated counts data from Broad firehose (tumor.uncv2.mRNAseq_RSEM_all.txt). Missing values due to the batch effect correction process were imputed using *K*-nearest neighbors (KNN). For each tumor type, values were imputed based on gene vectors under the assumption that genes are similarly expressed between samples of the same tumor type. For genes with less than 60% of samples with missing values, values were imputed using the logarithmic (base 2) of the gene expression value plus one, and subsequently back-transformed when writing the imputed matrices to file. The resulting gene expression matrix used for the Boruta analysis described below consisted of 870 TCGA samples and 16,309 protein-coding genes (based on "hgnc_complete_set.txt" downloaded from HGNC on 7.24.2018). To generate a RSEM raw counts matrix for the DESeq2 analysis described below, mRNAseq_Preprocess.Level_3 data was downloaded from Broad Firehose (tumor.uncv2.mRNAseq_raw_counts.txt).

Boruta analysis

To identify a minimal set of genes whose expression values were predictive of the sample being ecDNA(+), we used Boruta[10], an automated feature selection algorithm that utilizes multiple iterations of the random forest classifier to determine the statistical significance of selected features. The algorithm is terminated when all features are categorized as "confirmed" or "rejected", or until the user-defined number of iterations is reached. In our modified version of the BorutaPy python package (6.21.2021; *https://github.com/scikit-learn-contrib/boruta_py*), we set the maximum number of iterations to 400, a stagnant count maximum of 5, and a tentative count minimum of 50. This translates to termination of Boruta if 400 iterations are reached, or if the tentative count (features that have yet to be "confirmed" or "rejected") falls to or below 50 and these tentative features remain tentative for 5 iterations.

While we use a standard implementation of Boruta, the method is briefly described here for expository purposes. In each iteration, *i*, within a single Boruta trial, the input is a gene expression matrix, *M*, of dimension *r* x c, where *r* is the number of samples and *c* is the number of tentative or confirmed features (i.e., genes). Boruta generates *c* shadow feature vectors by random shuffling of feature vectors in matrix *M*, generating a new matrix *M'* of dimension *r* x 2*c*. A random forest classifier (class_weight=balanced_subsample, max_depth=7) is then used to quantify the importance of each feature in separating ecDNA(+) from ecDNA(-) samples. Specifically, there are two possible outcomes for a feature: 1) if the feature scores higher than the best scoring shadow feature, the feature is considered a "hit", and 2) if the feature scores lower than the best scoring shadow feature, the feature is considered a "non-hit". Features are rejected after *i* iterations, if the number of hits is not significantly higher than expected by chance, using a Bonferroni corrected *p*-value.

In order to evaluate the ability of selected features in predicting the ecDNA status of a tumor sample, we left out 20% of ecDNA(+) and 20% of ecDNA(-) samples for the hold-out testing dataset in the evaluation procedure described below, and performed Boruta on the gene expression matrix consisting of the remaining 80% of samples. However, due to the unequal representation of ecDNA(+) and ecDNA(-) samples within each of the tumor subtypes, we opted to generate 200 training (80%) and testing (20%) datasets to decrease the bias that may be introduced during random sampling. For each of the 200 datasets, a Boruta analysis was performed on the 80% training data. Features categorized as "confirmed" were considered as Boruta genes for that specific trial. Of the 941 Boruta genes combined across the 200 trials, 408 genes were present in at least 10 of the 200 Boruta trials, and subsequently defined as the Core set of genes in downstream analyses.

Highly co-expressed genes

To identify genes co-expressed with the core set of Boruta genes, hierarchical clustering of the 16,309 genes was performed using the R package pvclust[11] (ver. 2.2-0; dist.method=correlation, method=ward.D2, nboot=1000). A total of 843 significant clusters (AU > 0.95) with at least 1 Boruta gene were selected, consisting of 1,375 genes. To obtain the final list of CorEx genes, we apply a minimal count of 10 trials for the gene or 10 trials for the cluster of genes seen in 200 Boruta trials. A cluster is determined to be seen in a Boruta trial if at least one of its members is selected in the trial. This results in 354 clusters, with a total number of 643 genes, of which 408 are Core genes.

Evaluation of CorEx genes

To evaluate a set of genes, *G*, as predictive of ecDNA presence in tumor samples, we performed cross-validation and hyper-parameter tuning on each of the 80% training datasets, and evaluated the final model on the corresponding hold-out 20% testing dataset using the scikit-learn package. Specifically, the gene expression matrix for cross-validation and hyper-parameter tuning consisted of *m* samples from the training dataset and *n* genes from set *G*. RandomizedSearchCV (n_iter=50, cv=5, scoring=f1) was first used to narrow down a wide range of hyper-parameters for the random forest classifier (RandomForestClassifier, class_weight='balanced_subsample'), and GridSearchCV (cv=StratifiedKFold(n_splits=5, shuffle=True), scoring=f1) was then used to test every combination of a smaller range of hyper-parameters given the best parameters from RandomizedSearchCV. The hyper-parameters tuned (initial values) include the number of trees in the forest (n_estimators: np.linspace(100, 2000, num = 10)), the maximum depth of the tree (max_depth: None, np.linspace(10, 100, num = 10)), the minimum number of samples required to split an internal node (min_samples_split: (2, 5, 10)), and the minimum number of samples required to be at a leaf node (min_samples_leaf: (1, 2, 4)). The best estimator from the GridSearchCV hyper-parameter tuning was then evaluated on the 20% testing dataset, where the gene expression matrix consisted of *m* samples from the testing dataset and *n* genes from set *G*. Performing this procedure on each of the 200 training/testing datasets resulted in 200 data points for each of the 3 metrics computed using sklearn.metrics: precision_score, recall_score, and average_precision_score (AUPR).

We performed this procedure on the following sets of genes, *G*: the 408 Core genes, 408 randomly selected genes, the 643 CorEx genes, 643 randomly selected genes, a set of 643 most differentially expressed genes based on the absolute log-fold change estimates from a conventional DE analysis using DESeq2[15] as described below, and the set of 3,012 significant genes from the GLM analysis described below.

Generalized linear model (GLM) analysis

We tested each of 16,309 genes independently in a separate logistic regression model using the glm() function in the R stats package (v4.2.0), and retained genes that were significant (*p*-value 0.01). Specifically, the model was defined as glm($y \sim g_j + t$, data = *M*, family = binomial(link = 'logit')), where *y* is the response vector where y_i=1 if sample $i \in \{1,...,870\}$ is ecDNA(+) and y_i =0 otherwise, g_j is the vector of expression values for gene $j \in \{1,...,16309\}$ in samples $i \in \{1,...,870\}$, *t* is the covariate vector representing the tumor subtypes of samples $i \in \{1,...,870\}$, and *M* is the data matrix containing values of gene expression, tumor subtype, and ecDNA status for all samples. The equation for the binomial logistic regression described above is formulated as $log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$, where *p* is the probability that the dependent variable *y* is 1, *X* are the independent variables, and β are the coefficients of the model. In this case, *k*=1 represents independent variable gene *j* and *k*=2 represents the tumor subtype covariate *t*. Of the 16,309 genes tested independently, 3,012 genes were significant at *p*-value<0.01.

Default DE analysis

We performed a default DESeq2[15] (R package, ver. 1.36.0) analysis to obtain shrunken maximum *a posteriori* (MAP) log-fold change estimates for effect size (i.e., LFC). Specifically, to obtain i) LFC effect size values per gene for integration with its Cliff's delta effect size value when determining if a gene is up- or down-regulated in ecDNA(+) samples, and ii) a list of *n* top-ranked genes by absolute value of the LFC (with application of an adjusted *p*-value <0.05 cutoff and LFC threshold of $log_2(1.1) = 0.13$) for use in comparison against genes selected as important in the prediction of ecDNA in samples. For comparisons against Core genes, *n* is set to 408, and for comparisons against CorEx genes, *n*, is set to 643.

To obtain the LFC effect size metric between ecDNA(+) vs. ecDNA(-) samples for each gene's expression, we fed as input to DESeq2 a matrix of raw RSEM estimated counts. To take into account batch effects, we included the center and platform information of samples, downloaded from synapse id syn4976363 (EB++GeneExpAnnotation.tsv), in the design of the DESeq object:

DESeq_object <-DESeqDataSetFromMatrix(countData = AC_0_4_9_TCGA_matrix, colData = coldata, design = ~batch+condition)

To compute results, the lfcThreshold was set to $log_2(1.1)$ for an accurate computation of *p*-values and the contrast set to c("condition", "ecDNA(+)", "ecDNA(-)") to obtain the logarithmic fold change of the form $\left(\frac{ecDNA(+)}{ecDNA(-)}\right)$. By setting the lfcThreshold, the null hypothesis tested is that $|LFC| \leq \theta$, where $\theta = (1.1)$, and the alternative hypothesis is that $|LFC| > \theta$. A $log_2(1.1)$ value is chosen as the minimal value/negligible effect size threshold as it represents a 10% fold-change, and anything below this fold-change would likely not be of biological interest[54]. The specific commands run are as follows:

DESeq_object$condition <-relevel(DESeq_object$condition, ref = "Non_ecDNA") DESeq_object <-DESeq(DESeq_object) results <-results(DESeq_object, lfcThreshold= log2(1.1), contrast=c("condition","ecDNA","Non_ecDNA"))

To obtain the shrunken MAP log-fold change estimates, we used the lfcShrink function provided in DESeq2, using the default apeglm method for the empirical Bayes shrinkage procedure[55]:

lfcShrink(DESeq_object, lfcThreshold= log2(1.1), coef="condition_ecDNA_vs_Non_ecDNA", type="apeglm")

Up- or Down-regulated genes in ecDNA(+) samples

To categorize genes as "up-" or "down-" regulated in ecDNA(+) samples, we integrated two effect size metrics, Cliff's delta (*d*)[16,17] and the DESeq2 shrunken MAPlog-fold change estimate (LFC[15]). Effect size is a measure of the magnitude of deviation from the null hypothesis, and unlike *p*-values, has the advantage of not being impacted by sample size[56]. This property is especially useful when comparing effect size values of a gene between tests where sample sizes differ. Comparing *p*-values between such tests would be invalid.

Cliff's delta, *d*, is a non-parametric measure of the separation between two distributions and ranges from −1 to 1. Given two distributions, $X = \{i_1, i_2, ..., i_m\}$ and $Y = \{j_1, j_2, ..., j_n\}$, comparisons are made between each of m values in X and n values in Y. Cliff's delta is computed as $d = \frac{\#(i>j)-\#(i<j)}{mn}$, where $\#(i > j)$ is the number of times a member of X is greater than a member of Y, and $\#(i < j)$ is the number of times a member of X is less than a member of Y[16]. A negative d indicates that values in Y tend to be higher than X, while a positive d indicates that values in X tend to be higher than Y. The magnitude of the effect size of Cliff's delta can be separated into four levels: $|d| < 0.147$ for negligible effects, $0.147 \leq |d| < 0.33$ for small effects, $0.33 \leq |d| < 0.474$ for medium effects, and $|d| \geq 0.474$ for large effects[56]. The python package used to compute Cliff's delta values can be accessed at *https://github.com/neilernst/cliffsDelta*. The input values used to compute Cliff's delta are log-transformed normalized gene expression values plus one as described in the "Gene expression datasets" section of methods.

The DESeq2 LFC is computed as described above. To allow integration of the Cliff's delta effect size with the DESeq2 LFC, we also separated the LFC values into four levels: $|LFC| < log_2(1.1)$ for negligible effects, $log_2(1.1) \leq |LFC| < log_2(1.5)$ for small effects, $log_2(1.5) \leq |LFC| < log_2(2)$ or medium effects, and $|LFC| \geq log_2(2)$ for large effects.

For each gene, *g*, whether the gene is up-regulated or down-regulated in ecDNA(+) samples is determined by the signage and magnitude of its effect sizes d_{gg} and LFC_{gg}. The initial criteria for $d_{gg}$and $LFC_{gg}$to be used as a determinant in the direction of a gene is for it to have a magnitude larger than that of a negligible effect. If the signage of $d_{gg}$and $LFC_{gg}$are both positive, gene *g* is considered up-regulated in ecDNA(+). If only a single value has a magnitude larger than the negligible effect threshold (e.g, $d_{gg} > 0$ and $LFC_{gg} = -0.1$), gene *g* is considered up-regulated in ecDNA(+). In the case of conflicting signages between the two values, the effect size with a larger magnitude takes precedence. For example, if $d_{gg} = -0.2$ and $LFC_{gg} = 0.847$, given that d_{gg}has a small negative effect and LFC_{gg}has a large positive effect, gene *g* is considered up-regulated in ecDNA(+) samples.

Tumor heatmap

A Cliff's delta effect size matrix representing 643 CorEx genes was generated to compare TCGA with tumor expression patterns. For each of the 11 tumor types with at least 10 ecDNA(+) and at least 10 ecDNA(-) samples, we re-computed Cliff's delta. Using a Fisher's exact test (fisher_exact function from the scipy.stats python package; alternative hypothesis: two-sided), we tested the null-hypothesis of whether the up- and down-directionality of CorEx genes in TCGA vs. each tumor were independent of each other. The contingency table is as below. The directionality of a gene (up or down) was based solely on the signage of the gene's Cliff's delta effect size value.

		Tumor	
		UP	DOWN
TCGA	UP	a	b
	DOWN	c	d

Tcga

Gene Ontology (GO) enrichment analysis

To identify Gene Ontology Biological Process (GOBP) terms that were enriched in either the set of down-regulated or up-regulated CorEx genes, we applied one-sided Fisher's exact tests (alternative="greater"; scipy.stats python package) on 2×2 contingency tables for each GOBP term. Specifically, in the contingency table below, N is the total number of genes in the universe (i.e., 16k for the number of genes measured in the RNAseq data), n is the number of DE genes (either up- or down-regulated in ecDNA(+) samples), m is the number of genes belonging to the GOBP term as defined by gene sets from MSigDB (c5.go.bp.v7.5.1.entrez), and k is the number of DE genes that belong to the GOBP term. The false discovery rate was controlled at 5% and adjusted *p*-values were computed using the Benjamini-Hochberg procedure (fdr correction from python statsmodels package). A final set of GOBP terms with adjusted *p*-value<0.05 was used for downstream analysis.

	DE	Non-DE
Inside GOBP term	k	m-k
Outside GOBP term	n-k	N+k-n-m

Clustering gene sets

To cluster enriched gene sets into categories for visualization purposes, Cohen's kappa coefficient (python sklearn cohen_kappa_score) was used to determine term-term "connectivity" (agreement of term-term pairs) – an approach described in the DAVID paper[22], Given a $r \times c$ binary matrix, where enriched GOBP terms are rows, CorEx genes are columns, and values are 1 if a CorEx gene is part of the GOBP term or 0 otherwise, kappa scores were computed between each pair of terms, where term $t_i \in t$ and $i = \{1,2,3,\ldots. n\}$: $Kappa\_Score(t_x, t_y)$, where $x \in i$ and y $\in i$.

Each term, t_i, formed an initial seeding group, g_i, where a term t_x is part of g_i if $Kappa\_Score(t_i, t_x) \geq score\ threshold$. If at least 50% of term-term pairs in g_i have a $Kappa\_Score \geq score\ threshold$, the initial seeding group g_i is retained for the next step. The second criteria ensures that terms within the same seeding group have strong interconnectivity. An iterative merging of seeding groups then follows: groups sharing p% or more members are merged. The representative term for each group was determined as the member with the highest interconnectivity score with other members of the group. After the automatic grouping process, a manual inspection leads to the merging of outliers or smaller groups into representative groups. We used a kappa score threshold of 0.5, and condition of p≥25% of shared members when merging for the down-regulated genes, and a kappa score threshold of 0.6 and condition of p≥50% of shared members when merging for the up-regulated genes.

DDR pathway genes

We hand-curated 88 genes for double-stranded break DNA damage repair pathways (a-EJ, HR, c-NHEJ, SSA) via an extensive literature search, and added an additional 51 genes from the following MSigDB (c5.go.bp.v2022.1.Hs.entrez.gmt) GO biological process terms: GO:0097680 (double strand

break repair via classical nonhomologous end joining), GO:0097681 (double strand break repair via alternative nonhomologous end joining), GO:1905168 (positive regulation of double strand break repair via homologous recombination), and GO:0045002 (double strand break repair via single strand annealing). This resulted in a final list of 129 genes. The directionality of the genes, as either up- or down-regulated in ecDNA(+) samples, is based on the full set of 1,440 samples, consisting of 243 ecDNA(+) and 1,197 ecDNA(-) samples representing 24 tumor types (**Table S12**; **Table S13**).

To test whether the number of genes passing our effect size thresholds for all genes 5,256 (UP) and 5,251 (DOWN) was significantly different from the up-/down-regulated genes implicated in each of the 4 pathways, we performed a Fisher's exact test (fisher_exact function from the scipy.stats python package) on the contingency table below, where *c* and *d* are the up- and down-regulated genes for each of the pathways tested.

Contingency table:

	UP	DOWN
All genes	5,256	5,251
Pathway	*c*	*d*

Pathway values:

	c	*d*
c-NHEJ	14	7
Alt-EJ	11	1
HR	46	8
SSA	11	1

Physical presence on amplicons

To determine physical presence of a gene on an amplicon, gene coordinates listed in the gene annotations file GRCh37/human_hg19_september_2011/Genes_July_2010_hg19.gff, downloaded from the AA repo on 3/21/2022, were mapped to amplicon genomic intervals (bed files). A gene is determined to be physically present on an amplicon if its genomic coordinates are fully encompassed within the amplicon genomic intervals.

Mutational analysis

We pulled the list of mutations from the open-access version of MC3 dataset (*https://ellrottlab.org/project/mc3/*), then investigated differences between ecDNA(+) and ecDNA(-) samples. Synonymous mutations were excluded when calculating the mutation burden.

For the differentially mutated gene analysis, only damaging mutations were selected by using snpEff[57] annotation which is originally included in the MC3 dataset. First, mutations annotated as HIGH in the IMPACT column were selected to obtain frameshift INDELs and stop gain SNVs. Next, mutations predicted to be damaging by SIFT[58] and PolyPhen2[59], were selected to obtain damaging missense mutations. Finally, we generated 2-by-2 contingency tables for each gene with cells *a*, *b*, *c*, and *d* representing the number of individuals with and without damaging mutations

in ecDNA(+) and (-) tumors. The odds ratios were computed as $OR = \frac{ad}{bc}$, where a is the number of individuals in ecDNA(+) with the mutation, b is the number of individuals in ecDNA(+) without the mutation, c is the number of individuals in ecDNA(-) with the mutation, and d is the number of individuals in ecDNA(-) without the mutation. To determine if a gene contained mutations that were implicated in cancer, we checked genes against the Cancer Gene Census (CGC) database (v97)[60], marking genes in the database as CGC and those that were not as non-CGC.

Classification with mutational status

First, we created a binary matrix representing whether a gene is damaged or not, from the MC3 damaging mutation set described as above. Then we divided the whole matrix into 80% of the training set and 20% of the test set. Hyperopt was applied on the training set to select the best parameters for the XGBoost[40] model. The optimal parameters estimated by Hyperopt (eta = 0.1, max_depth = 6, min_child_weight = 3.0, scale_pos_weight = 4.9) were input to the XGBoost model, and the ecDNA status of each sample were also input as an answer set. The performance of the model was checked by inputting the 20% test set to the model and comparing the output result with the answer. Principal component analysis was also performed with the same mutational matrix as above, using the scikit.learn package.

Ranking of CorEx genes

To rank CorEx genes by importance, we computed harmonic mean rank values based on three categories: a) the average Gini importance statistic or mean decrease impurity (MDI) MDI (feature importance) values extracted from the trained random forest models on 200 training sets during the evaluation method described above, b) the number of Boruta trials (out of 200) that a gene is selected in, rounded to 2-digits, and c) the number of Boruta trials (out of 200) that a gene is selected in when counting by cluster, rounded to 2-digits. The ranks for each category are adjusted separately so that genes with the same value share the same rank value. For example, if using the Boruta trial count as the rank value:

	Trial count	round(trial_count/10.0)	Rank	Adjusted rank
Gene A	200	20	1	1
Gene B	200	20	2	1
Gene C	200	20	3	1
Gene D	187	19	4	4

The harmonic mean rank is defined as:

$$harmonic\ mean\ rank = \frac{3}{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}}$$

Tumor immune subtype

We classified ecDNA(+) and ecDNA(-) samples into the 6 immune subtype categories (Thorsson et al., 2018[38]) provided in Table S6 from Bagaev et al., 2021[61].

Impact of tumor purity on CorEx gene expression

To investigate the effects of the presence of non-cancer tissue (impurity) in bulk RNA-seq samples on the analyses performed in this study, we utilized the consensus measurement of purity estimations (CPE) for TCGA samples from a publication by Aran et al.[14]. Of the 870 TCGA samples (234 ecDNA(+), 636 ecDNA(-)) with gene expression (RNA-seq) data, 701 samples (174 ecDNA(+), 527 ecDNA(-)) were assigned a CPE value by Aran et al.. To determine if the presence of undetected ecDNA in ecDNA(-) samples would confound the results by reducing the discriminating power of genes, we measured the expression directionality of CorEx genes in all samples (n=870) versus samples which had a high tumor purity (CPE≥0.8, n=287). Specifically, *p*-values were obtained by performing Mann-Whitney U rank tests (scipy.stats python package) on gene expression values of ecDNA(+) and ecDNA(-) samples for both the 870 TCGA samples and 287 TCGA samples with high tumor purity. Genes with a significantly (*p*-value≤0.05) higher expression in ecDNA(+) samples (alternative="greater") were labeled as "UP", while genes with a significantly lower expression in ecDNA(+) samples (alternative= "less") were labeled as "DOWN". To generate a plot that compared gene directionality of all samples vs. high purity samples using *p*-values, a function F was applied to *p*-values. Specifically, $F(p) = d \cdot log_{10}(p)$, where p is the *p*-value and $d = 1$ if directionality is "DOWN" and $d = -1$ if directionality is "UP".

Author Contributions

M.S.L., P.S.M, and V.B. conceived and designed the experiments and core narrative of the manuscript. All analyses were performed by M.S.L., apart from the following ones. Mutational analysis was performed by S.J. The amplicon classifications were performed by J.L. S.W, V.B, H.Y.C. and P.S.M. provided guidance on the analysis and interpretation of results. M.S.L., S.J., J.L., and V.B. wrote the manuscript with feedback from all authors.

Acknowledgements

This project was supported by Cancer Grand Challenges CGCSDF-2021\100007 with support from Cancer Research UK and the National Cancer Institute (V.B, H.Y.C., P.S.M.) M.S.L., J.L., and V.B. were supported in part by grants U24CA264379, R01GM114362, and OT2CA278635 from the NIH and CGCATF-2021/100025 from CRUK. S.J. was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI19C1330).

Competing Interests

V.B. is a co-founder, paid consultant, SAB member and has equity interest in Boundless Bio, inc. and Abterra Biosciences, Inc. H.Y.C. is a co-founder of Accent Therapeutics, Boundless Bio, Cartography Biosciences, Orbital Therapeutics, and an advisor of 10x Genomics, Arsenal Biosciences, Chroma Medicine, and Spring Discovery. P.S.M. is a co-founder and advisor of Boundless Bio. J.L. receives compensation as a consultant for Boundless Bio. The remaining authors declare no competing interests.

References

1. Turner K. M., et al. (2017) **Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity** *Nature* **543**:122–125

2. Kim H., et al. (2020) **Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers** *Nat. Genet* **52**:891–897

3. Nathanson D. A., et al. (2014) **Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA** *Science* **343**:72–76

4. Lange J. T., et al. (2022) **The evolutionary dynamics of extrachromosomal DNA in human cancers** *Nat. Genet* **54**:1527–1533

5. Luebeck J., et al. (2023) **Extrachromosomal DNA in the cancerous transformation of Barrett's oesophagus** *Nature* https://doi.org/10.1038/s41586-023-05937-5

6. Hung K. L., et al. (2021) **ecDNA hubs drive cooperative intermolecular oncogene expression** *Nature* **600**:731–736

7. Wu S., et al. (2019) **Circular ecDNA promotes accessible chromatin and high oncogene expression** *Nature* **575**:699–703

8. Morton A. R., et al. (2019) **Functional Enhancers Shape Extrachromosomal Oncogene Amplifications** *Cell* **179**:1330–1341

9. van Leen E., Brückner L., Henssen A. G (2022) **The genomic and spatial mobility of extrachromosomal DNA and its implications for cancer therapy** *Nat. Genet* **54**:107–114

10. Kursa M. B., Rudnicki W. R (2010) **Feature Selection with the Boruta Package** *J. Stat. Softw* **36**

11. Suzuki R., Shimodaira H (2006) **Pvclust: an R package for assessing the uncertainty in hierarchical clustering** *Bioinformatics* **22**:1540–1542

12. Kobayashi Y., et al. (2021) **Mitotic checkpoint regulator RAE1 promotes tumor growth in colorectal cancer** *Cancer Sci* **112**:3173–3189

13. Wang Q., et al. (2019) **Prognostic Potential of Alternative Splicing Markers in Endometrial Cancer** *Mol. Ther. - Nucleic Acids* **18**:1039–1048

14. Aran D., Sirota M., Butte A. J (2015) **Systematic pan-cancer analysis of tumour purity** *Nat. Commun* **6**

15. Love M. I., Huber W., Anders S (2014) **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2** *Genome Biol* **15**

16. Cliff N (1993) **Dominance statistics: Ordinal analyses to answer ordinal questions** *Psychol. Bull* **114**:494–509

17. Cliff N (1996) **Answering Ordinal Questions with Ordinal Data Using Ordinal Statistics** *Multivar. Behav. Res* **31**:331–350

18. Garsed D. W., et al. (2014) **The Architecture and Evolution of Cancer Neochromosomes** *Cancer Cell* **26**:653–667

19. Subramanian A., et al. (2005) **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles** *Proc. Natl. Acad. Sci* **102**:15545–15550

20. Liberzon A., et al. (2011) **Molecular signatures database (MSigDB) 3.0** *Bioinformatics* **27**:1739–1740

21. Liberzon A., et al. (2015) **The Molecular Signatures Database Hallmark Gene Set Collection** *Cell Syst* **1**:417–425

22. Huang D., et al. (2007) **The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists** *Genome Biol* **8**

23. Feng Y., et al. (2021) **Homeobox Genes in Cancers: From Carcinogenesis to Recent Therapeutic Intervention** *Front. Oncol* **11**

24. Fish L., et al. (2021) **A prometastatic splicing program regulated by SNRPA1 interactions with structured RNA elements** *Science* **372**

25. Cui H., et al. (2022) **TNFAIP6 Promotes Gastric Carcinoma Cell Invasion via Upregulating PTX3 and Activating the Wnt/β-Catenin Signaling Pathway** *Contrast Media Mol. Imaging 2022* :1–9

26. Li Y., et al. (2019) **TMEM203 is a binding partner and regulator of STING-mediated inflammatory signaling in macrophages** *Proc. Natl. Acad. Sci* **116**:16479–16488

27. Feng Z., et al. (2022) **RNF114 Silencing Inhibits the Proliferation and Metastasis of Gastric Cancer** *J. Cancer* **13**:565–578

28. Ribeiro J. R., Lovasco L. A., Vanderhyden B. C., Freiman R. N (2014) **Targeting TBP-Associated Factors in Ovarian Cancer** *Front. Oncol* **4**

29. Xu N., et al. (2020) **SHCBP1 promotes tumor cell proliferation, migration, and invasion, and is associated with poor prostate cancer prognosis** *J. Cancer Res. Clin. Oncol* **146**:1953–1969

30. Chang H. H. Y., Pannunzio N. R., Adachi N., Lieber M. R (2017) **Non-homologous DNA end joining and alternative pathways to double-strand break repair** *Nat. Rev. Mol. Cell Biol* **18**:495–506

31. Ghosh D., Raghavan S. C (2021) **20 years of DNA Polymerase μ, the polymerase that still surprises** *FEBS J* **288**:7230–7242

32. Daley J. M., Sung P (2014) **53BP1, BRCA1, and the Choice between Recombination and End Joining at DNA Double-Strand Breaks** *Mol. Cell. Biol* **34**:1380–1388

33. Fouquin A., et al. (2017) **PARP2 controls double-strand break repair pathway choice by limiting 53BP1 accumulation at DNA damage sites and promoting end-resection** *Nucleic Acids Res* **45**:12325–12339

34. He L., Guo Z., Wang W., Tian S., Lin R (2023) **FUT2 inhibits the EMT and metastasis of colorectal cancer by increasing LRP1 fucosylation** *Cell Commun. Signal. CCS* **21**

35. Lawrence T (2009) **The Nuclear Factor NF-B Pathway in Inflammation** *Cold Spring Harb. Perspect. Biol* **1**:a001651–a001651

36. Urban-Wojciuk Z., et al. (2019) **The Role of TLRs in Anti-cancer Immunity and Tumor Rejection** *Front. Immunol* **10**

37. Chen D. S., Mellman I (2013) **Oncology Meets Immunology: The Cancer-Immunity Cycle** *Immunity* **39**:1–10

38. Thorsson V., et al. (2018) **The Immune Landscape of Cancer** *Immunity* **48**:812–830

39. Wu T., et al. (2022) **Extrachromosomal DNA formation enables tumor immune escape potentially through regulating antigen presentation gene expression** *Sci. Rep* **12**

40. Chen T., Guestrin C., Association for Computing Machinery (2016) **XGBoost: A Scalable Tree Boosting System** *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794* https://doi.org/10.1145/2939672.2939785

41. Bergstrom E. N., et al. (2022) **Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA** *Nature* **602**:510–517

42. Hadi K., et al. (2020) **Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs** *Cell* **183**:197–210

43. Luca M. D., Lavia P., Guarguaglini G (2006) **A Functional Interplay Between Aurora-A, Plk1 and TPX2 at Spindle Poles: Plk1 Controls Centrosomal Localization of Aurora-A and TPX2 Spindle Association** *Cell Cycle* **5**:296–303

44. Guo Y., Shi J., Zhao Z., Wang M (2021) **Multidimensional Analysis of the Role of Charged Multivesicular Body Protein 7 in Pan-Cancer** *Int. J. Gen. Med* **14**:7907–7923

45. Innes A. J., et al. (2021) **XPO7 is a tumor suppressor regulating p21 CIP1-dependent senescence** *Genes Dev* **35**:379–391

46. Qin B., et al. (2015) **DBC1 functions as a tumor suppressor by regulating p53 stability** *Cell Rep* **10**:1324–1334

47. Von Hoff D. D., et al. (1992) **Elimination of extrachromosomally amplified MYC genes from human tumor cells reduces their tumorigenicity** *Proc. Natl. Acad. Sci* **89**:8165–8169

48. Gupta R. A., et al. (2010) **Long noncoding RNA HOTAIR reprograms chromatin state to promote cancer metastasis** *Nature* **464**:1071–1076

49. Sanborn J. Z., et al. (2013) **Double Minute Chromosomes in Glioblastoma Multiforme Are Revealed by Precise Reconstruction of Oncogenic Amplicons** *Cancer Res* **73**:6036–6045

50. Samson N., Ablasser A. (2022) **The cGAS–STING pathway and cancer** *Nat. Cancer* https://doi.org/10.1038/s43018-022-00468-w

51. Sun L., Wu J., Du F., Chen X., Chen Z. J (2013) **Cyclic GMP-AMP Synthase Is a Cytosolic DNA Sensor That Activates the Type I Interferon Pathway** *Science* **339**:786–791

52. Rizvi N. A., et al. (2015) **Mutational landscape determines sensitivity to PD-1 blockade in non– small cell lung cancer** *Science* **348**:124–128

53. McGrail D. J., et al. (2021) **High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types** *Ann. Oncol* **32**:661–672

54. McCarthy D. J., Smyth G. K (2009) **Testing significance relative to a fold-change threshold is a TREAT** *Bioinformatics* **25**:765–771

55. Zhu A., Ibrahim J. G., Love M. I (2019) **Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences** *Bioinformatics* **35**:2084–2092

56. Romano Jeanine, Kromrey Jeffrey D., Coraggio Jesse, Skowronek Jeff (2006) **Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen'sd for evaluating group differences on the NSSE and other surveys?** *Annu. Meet. Fla. Assoc. Institutional Res* **177**

57. Cingolani P., et al. (2012) **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3** *Fly (Austin)* **6**:80–92

58. Ng P. C., Henikoff S (2003) **SIFT: Predicting amino acid changes that affect protein function** *Nucleic Acids Res* **31**:3812–3814

59. Adzhubei I., Jordan D. M., Sunyaev S. R, Unit7.20 (2013) **Predicting functional effect of human missense mutations using PolyPhen-2** *Curr. Protoc. Hum. Genet* **7**

60. Sondka Z., et al. (2018) **The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers** *Nat. Rev. Cancer* **18**:696–705

61. Bagaev A., et al. (2021) **Conserved pan-cancer microenvironment subtypes predict response to immunotherapy** *Cancer Cell* **39**:845–865

Article and author information

Miin S. Lin
Bioinformatics and Systems Biology Graduate Program, University of California at San Diego, La Jolla, CA, USA
ORCID iD: 0000-0003-2017-4246

Se-Young Jo
Department of Biomedical Systems Informatics and Brain Korea 21 PLUS Project for Medical Science, Yonsei University College of Medicine, Seoul, South Korea

Jens Luebeck
Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA, USA

Howard Y. Chang
Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA, USA, Department of Genetics, Stanford University, Stanford, CA, USA, Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA

Sihan Wu
Children's Medical Center Research Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA

Paul S. Mischel
Sarafan Chemistry, Engineering, and Medicine for Human Health (Sarafan ChEM-H), Stanford University, Stanford, CA, USA, Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA
For correspondence: pmischel@stanford.edu

Vineet Bafna
Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA, USA, Halıcıoğlu Data Science Institute, University of California at San Diego, La Jolla, CA, USA
For correspondence: vbafna@ucsd.edu
ORCID iD: 0000-0002-5810-6241

Copyright

Editors

Reviewing Editor
Thomas Gingeras
Cold Spring Harbor Laboratory, Cold Spring Harbor, United States of America

Senior Editor
Richard White
Ludwig Institute for Cancer Research, University of Oxford, Oxford, United Kingdom

Reviewer #1 (Public Review):

Recently discovered extrachromosomal DNA (ecDNA) provides an alternative non-chromosomal means for oncogene amplification and a potent substrate for selective evolution of tumors. The current work aims to identify key genes whose expression distinguishes ecDNA+ and ecDNA- tumors and the associated processes to shed light on the biological mechanisms underlying ecDNA genesis and their oncogenic effects. This is clearly an important question and through detailed analysis this work points to specific GO processes associated (up and down) with ecDNA+ tumors, namely, specific DNA damage repair processes and specific oncogenic processes.

In the initial submission I had commented on lack of clarity of method, potential biases, and in some cases inappropriate interpretation. In the revised version, the authors have addressed all my comments satisfactorily and I think this is an important work furthering our understanding of mechanisms underlying ecDNA+ tumors.

https://doi.org/10.7554/eLife.88895.2.sa2

Reviewer #2 (Public Review):

In their manuscript Lin et al. describe an important study on the transcriptional programs associated with the presence of extrachromosomal DNA in a cohort of 870 cancers of different origins. The authors find that compared to cancers lacking such amplifications, ecDNA+ cancers express higher levels of DNA damage repair-associated genes, but lower levels of immune-related gene programs.

This work is very timely and its findings have the potential to be very impactful, as the transcriptional context differences between ecDNA+ and ecDNA- cancers are currently largely unknown. The observation that immune programs are downregulated in ecDNA+ cancers may initiate new preclinical and translational studies that impact the way ecDNA+ cancers are treated in the future. Thus, this study has important theoretical implications that have the potential to substantially advance our understanding of ecDNA+ cancers.

Strengths:

The authors provide compelling evidence for their conclusions based on large patient datasets. The methods they used and analyses are rigorous.

Weaknesses:

The biological interpretation of the data remains observational. The direct implication of these genes in ecDNA(+) tumors is not tested experimentally.

https://doi.org/10.7554/eLife.88895.2.sa1

Reviewer #3 (Public Review):

Summary:

Using a combination of approaches, including automated feature selection and hierarchical clustering, the author identified a set of genes persistently associated with extrachromosomal DNA (ecDNA) presence across cancer types. The authors further validated the gene set identified using gene ontology enrichment analysis and identified that upregulated genes in extrachromosomal DNA-containing tumors are enriched in biological processes like DNA damage and cell proliferation, whereas downregulated genes are enriched in immune response processes.

Comments for the previous version:

Major comments:

(1) The authors presented a solid comparative analysis of ecDNA-containing and ecDNA-free tumors. An established automated feature selection approach, Boruta, was used to select differentially expressed genes (DEG) in ecDNA(+) and ecDNA(-) TCGA tumor samples, and the iterative selection process and two-tier multiple hypothesis testing ensured the selection of reliable DEGs. The author showed that the DEG selected using Boruta has stronger predictive power than genes with top log-fold changes.

(2) The author performed a thorough interpretation of the findings with GO enrichment analysis of biological processes enriched in the identified DEG set and presented interesting findings, including the enrichment in DNA damage process among the genes upregulated in ecDNA(+) tumors.

(3) Overall, the authors achieved their aims with solid data mining and analysis approaches applied to public data tumor data sets.

(4) While it may not be the scope of this study, it will be interesting to at least have some justification for choosing Boruta over other feature selection methods, such as Recursive Feature Elimination (RFE) and backward stepwise selection.

(5) The authors showed that DESEQ-selected DEGs with top log-fold changes have less strong predictive power and speculated that this may be due to the fact that genes with top log-fold changes (LFC) are confined only to a small subset of samples. It will be interesting to select DEGs with top log-fold changes after first partitioning the tumor samples. For example, randomly partition the tumor samples, identify the DEGs with top LFC, combine the DEGs identified from each partition, then evaluate the predictive power of these DEGs against the Boruta-selected DEGs.

(6) While the authors showed that the presence of mutations was not able to classify ecDNA(+) and (-) tumor samples, it will be interesting to see if variant allele frequencies of the genes containing these mutations have predictive power.

Comments for the revised version:

The authors addressed the comments and recommendations with solid analysis and explanations in the revision. The added analysis using GLM is especially appreciated and provides convincing evidence for the predicting power of the Boruta-selected genes. The only comment is at this point is that it is recommended that the author provide some justification for choosing Boruta over other feature selection methods. It is not necessary to provide benchmarking results - justification based on the review of previous literature is sufficient, as it is not well explained in the paper why Boruta was chosen in the first place. Is it state-of-the-art? Has it demonstrated better performance in other settings? A few sentences answering these questions should suffice.

https://doi.org/10.7554/eLife.88895.2.sa0

Author Response

The following is the authors' response to the original reviews.

> ***eLife assessment***
>
> *This study of extrachromosomal DNA (ecDNA) aims to identify genes that distinguish ecDNA+ and ecDNA- tumors. This timely study is important in addressing the genes responding to the amplification of the ecDNA. The data presented are for the most part solid, there were concerns regarding the clarity in the description of the analysis methods and whether the evidence for specific genes required to maintain the ecDNA+ state was entirely conclusive.*
>
> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Recently discovered extrachromosomal DNA (ecDNA) provides an alternative non-chromosomal means for oncogene amplification and a potent substrate for selective evolution of tumors. The current work aims to identify key genes whose expression distinguishes ecDNA+ and ecDNA- tumors and the associated processes to shed light on the biological mechanisms underlying ecDNA genesis and their oncogenic effects. While this is clearly an important question, the analysis and the evidence supporting the claims*

> *are weak. The specific machine learning approach seems unnecessarily convoluted, insufficiently justified and explained, and the language used by the authors conflates correlation with causality. This work points to specific GO processes associated (up and down) with ecDNA+ tumors, many of which are expected but some seem intriguing, such as association with DSB pathways. My specific comments are listed below.*

Response. As some of the specific questions below address similar concerns, we have answered them briefly here. As a high level point, the reviewer is correct in that other statistical or ML approaches could potentially have been used, and that some are simpler. However, the test used here directly addresses the question: Find a collection of genes whose expression value is predictive of ecDNA status in the sample. Because the underlying method in the Boruta analysis uses random forests, it can test predictive power without relying on a linearity assumption implicit in other methods. In this revision, we also compare against a Generalized Linear Model and show that it is less suited to the specific task above. We also address the reviewer concerns about specific parameter choices by showing robustness to the specific parameter.

> *(A) The claim of identifying genes required to 'maintain' ecDNA+ status is not justified - predictive features are not necessarily causal.*

Response. We agree with the reviewer that predictive features are correlative and not causal. In the manuscript, we identify genes whose expression (when used as a feature) is predictive of ecDNA presence or absence. Such predictive genes are consistently over-expressed or consistently under-expressed in ecDNA(+) samples relative to ecDNA(-) samples even though they are not required to be on ecDNA. To our knowledge, we did not claim that these genes are causal for ecDNA formation or maintenance, only that such genes and the underlying biological processes are worth investigating. In the beginning of the manuscript, we had written the following paragraph, but we have removed the last line (struck out here):

“In lieu of identifying genes that are highly differentially expressed between ecDNA(+) and ecDNA(-) samples but driven by a small subset of cases (e.g. gene A in Fig. S1a), we sought to identify genes (e.g. gene B) whose expression level was predictive of ecDNA presence. We assumed that genes that were persistently over-expressed or under-expressed in ecDNA(+) samples relative to ecDNA(-) samples were more likely to be involved in ecDNA biogenesis or maintenance, or in mediating the cellular response to the presence of ecDNA.”

We revised the manuscript to make sure that there are no claims that refer to causality. We revisited all phrases where the words like “maintain” were used and added appropriate disclaimers, or replaced them by the phrase, “ecDNA presence.” The remaining statements say, for example, “These results are consistent with a pan-cancer role of CorEx genes in ecDNA biogenesis and maintenance,” and do not claim causality.

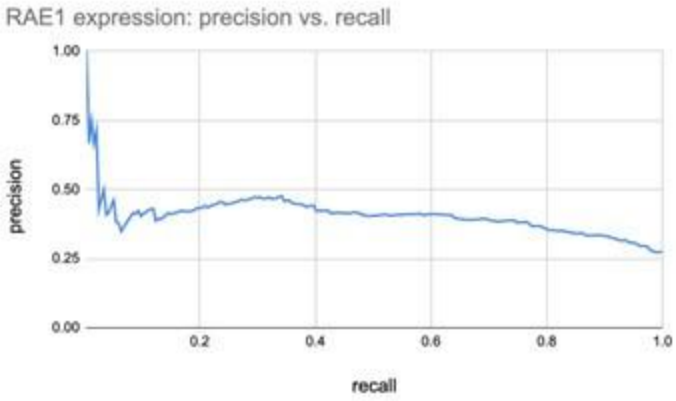
> *(B) The methods and procedures to identify the key genes is hyper-parameterized and convoluted and casts doubt on the robustness of the findings given the size and heterogeneity of the data.*
>
> *(a) In the first two paragraphs of Boruta Analysis Methods section, authors describe an iterative procedure where in each iteration, a binomial p-value is computed for each gene based on number of iterations thus far in which the gene was selected (higher GINI index than max of shadow features). But then in the third paragraph they simply perform Random Forest in 200 random 80% of samples and pick a gene if it is selected in at least 10/200. It is ultimately not clear what was done. Why 10/200? Also "the probability that a gene is a "hit" or "non-hit" in each iteration is 0.5" is unclear. That probability is of a gene achieving GINI index higher than the max of shadow features. How can it be 0.5?*

Response. We believe that there is some misunderstanding about the algorithm, and we agree that the description should have been more clear. We have greatly simplified the description in the manuscript. However, we want to provide some higher-level explanation here. Boruta is a standard feature extraction algorithm (Kursa, Journal of Statistical Software September 2010, Volume 36, Issue 11), and we used a Python implementation of the method. Given a gene expression data-set with class labels on samples, Boruta extracts features (genes) that best predict the class labels using a Random Forest Classifier, as long as the features are more predictive than permuted features added in each iteration. As we are using an implementation of a published method, we have removed non-essential details, referring directly to the publication. Nevertheless, to address the reviewer's specific critique, the number of false-features added changes in each iteration (it equals the number of accepted+uncommitted features). Therefore, the choice of 0.5 by Boruta (it is fixed in the published method and not a user-specified parameter) is a conservative approach. If a gene was no better than a randomly chosen feature, its predictive performance would exceed the most predictive randomly chosen feature by at most 0.5 (but could be lower, making the choice of 0.5 conservative).

While Boruta iteratively picks genes that are significantly better than random features, the list of genes predicted might be specific to the data-set, and might change with different data-sets. Therefore, we employed a bootstrapping strategy: we performed 200 trials each time picking 80% of the ecDNA(+) samples and 80% of the ecDNA(-) samples at random, thus generating many data-sets while maintaining class imbalance. For each of the 200 trials, we performed a Boruta analysis. Finally, we picked a gene if it was selected as a Boruta feature in at least 10 of 200 trials.

The reviewer has a reasonable critique about why 10 (of 200) specifically, and why not fewer or more. Most genes are weak predictors by themselves. For example, RAE1, which is the top ranked gene, picked in all 200 Boruta trials, can only predict ecDNA status with poor recall for any meaningful precision.

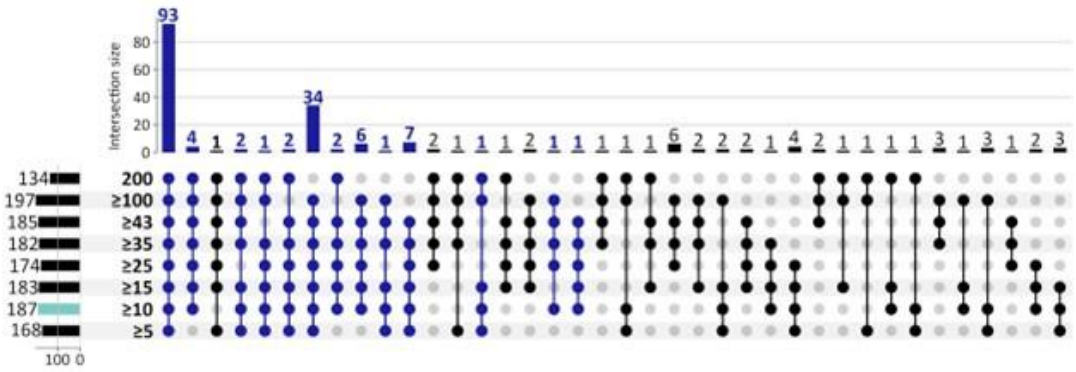
Author response image 1.



Given the weakness of an individual gene as a classifier, its repeated selection in multiple Boruta trials is already a significant event. By requiring a gene to be picked in 5% of the trials (10/200), we were selecting a small, but more robust list of genes. However, to further explore the reviewer's concerns, we also applied 8 other selection criteria ranging from 5 (of 200 Boruta trials) to 200 of 200 Boruta trials. See Figure below. The number of CorEx genes expectedly decreases. However, of the 187 GO terms that were enriched by 262 UP-genes using 10 of 200 Boruta trials as the selection criteria, 93 terms (49.7%) were enriched for each

cut-off (see Author response image 2), and 155 terms (82.9%) were enriched in at least 5 of the 8 cut-off criteria. Given that the remaining analysis works on the hierarchy of GO terms and finds 4 GO-categories (Mitotic Cell Cycle, G1/S, G2/M; cell-division; DSB DNA Damage response; and the HOX Gene cluster) enriched by UP-regulated genes, those conclusions would hold regardless of the specific cut-off.

Author response image 2.



The number of GO terms that were enriched by DOWN-regulated genes is smaller, only 73, and falls rapidly for higher cut-offs, with 25 at a cut-off of 15. Therefore we see fewer terms enriched for more stringent cut-offs. However, they all support immune processes. These results do suggest that there are fewer genes that are consistently down-regulated in ecDNA(+) cancers, and expression change in a small number of genes may be sufficient to promote conditions for ecDNA.

Finally, we note that in the final section we discuss the 65 most highly ranked genes with a harmonic mean rank <= 3. These 65 CorEx genes (or a member of their cluster) appear in each of 200 Boruta trials. Thus, their choice is also not dependent on the cut-off of 10 in 200. In summary, the conclusions of the paper do not depend upon the specific cut-off of 10 in 200 trials.

We have added the figure as a supplemental figure and have added the following text to the manuscript on pages 17 and 18.

“Any CorEx gene is either a Core gene that was selected as a feature in at least 5% of 200 Boruta trials, or be highly co-expressed with a Core gene. Because the selection criterion of 5% is arbitrary, we also tested robustness with 8 other cut-offs ranging from 5-of-200 to 200-of-200 Boruta trials. The number of CorEx genes expectedly decreases with more stringent cut-offs. However, of the 187 GO terms that were enriched by 262 CorEx UP-genes using 10 of 200 Boruta trials as the selection criteria, 93 terms (49.7%) were enriched for each cut-off (Fig. S9), and 155 terms (82.9%) were enriched in at least 5 of the 8 cut-offs. Given that our subsequent analyses utilized the hierarchy of GO terms and identified 4 GO-categories enriched by UP-regulated genes, the conclusions would hold regardless of the specific cut-off.”

> *(b) The approach of combining genes with clusters is arbitrary. Why not start with clusters and evaluate each cluster (using some gene set summary score) for their ability to discriminate? Ultimately, one needs additional information to disambiguate correlated genes (i.e. in a coexpression cluster) in terms of causality.*

Response. In general, the approach proposed by the reviewer is reasonable. However, we did consider that possibility and found that our approach was easier to implement. For example,

if we clustered first, we would have the challenge of choosing the correct set of clusters. Also, the Boruta analysis would become very difficult while dealing with clusters (e.g., how to define falsefeatures?). We tested other methods of picking genes that were suggested by other reviewers such as generalized linear models. They turned out not to be as predictive of ecDNA status, as described later in the response. Finally, we performed many experiments to ensure the validity of the clustering. Specifically, we had the following text in the paper:

“Notably, among the 354 clusters, only 2 clusters (with 14 total genes) did not contain any Core genes. As most genes do not have completely identical expression patterns, we would expect one gene to be consistently picked as a Boruta gene over another co-expressed gene. Consistent with this hypothesis, most (344/354) clusters contained only 1 or 2 Core genes (Fig. 1c). When selecting clusters that contained at least 1 Core and 1 co-expressed gene, 53 of 71 clusters contained 1 to 3 Core genes (Fig. S1b), confirming that a few genes per co-expressed cluster provide sufficient predictive value, but other co-expressed genes might still play an important functional role in maintaining ecDNA(+) status.”

These experiments suggest that the genes found by extending the Core genes through clustering do not radically change the Core genes, but only enhance the set.

> *(c) The cross-validation procedure is not clear at all. There is a mention of 80-20 split but exactly how/if the evaluation is done on the 20% is muddled. The way precision-recall procedure is also a bit convoluted - why not simply use the area under the PR curve?*

Response. We apologize if the method was unclear. We have rewritten the methods part to make things clearer. As a high level point, there are two places where we use the same 80-20 split, and that resulted in some confusion. We start by randomly picking 80% of the ecDNA(+) and 80% of ecDNA(-) samples to create an 80-20 split of all samples. This procedure is repeated to generate 200 80-20 split data-sets. These data-sets are hereafter called 200 training and test samples.

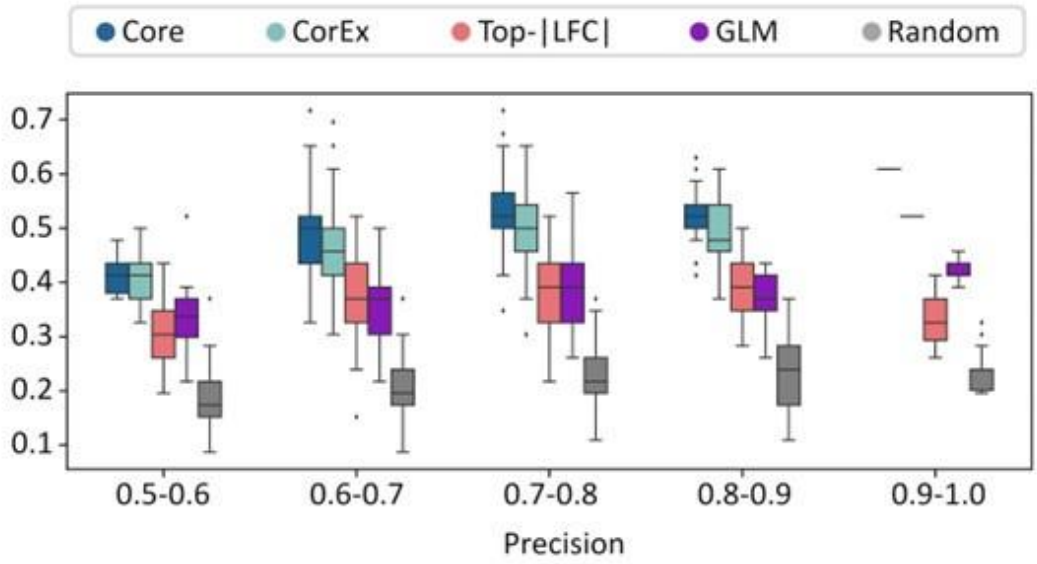
In the first usage, we use only the ‘training’ part of the 200 samples. We apply Boruta to each training set, and this helps us select the Core genes, which are then expanded to form the CorEx set. At this point, the CorEx genes are frozen for analysis in the rest of the paper. One question that we subsequently answer is what is the predictive power of the CorEx genes in determining if the sample is ecDNA(+) or ecDNA(-)? We also compare the predictive performance of CorEx genes relative to (a) Core genes, (b) LFC genes, and (c) random genes. In the revised manuscript, we have added another list of 3,012 genes selected using a single gene generalized linear model (GLM) for feature prediction. To make these comparisons, we utilized the same 200 training and test data-sets as before. In each test, we trained a random forest classifier on the training set and predicted on the ‘test’ set, for each of the 5 gene lists. This provided a uniform and fair method for testing which of the 5 gene lists was the better predictor of ecDNA status.

The precision recall values are plotted in Fig. 2b (also included below). We note that none of the gene lists was a great predictor of ecDNA status of a sample. However, the CorEx and Core genes were significantly more predictive than GLM, LFC, and random genes. The predictive power of GLM genes was very similar to LFC, and better than random.

For each of these 200 tests, we obtained a separate area under the precision-recall curve number for each of the gene-sets. To address the reviewer’s comments regarding a single number, we reported the average of the AUPRC for each of the gene-sets in the revision. The mean AUPRC values were added to the manuscript and are described here as well: Core_408_genes: 0.495 CorEx_643_genes: 0.48 Random_643_genes: 0.36 top_lfc_643_genes: 0.429 GLM_R_3012_genes: 0.426

We also changed Figure 2b to show box-plots showing distribution of recall values for specific precision windows instead of maximum recall. For ease of checking, the figure is reproduced below.

Author response image 3.



> *(d) The claim is that Boruta genes are different from differentially expressed genes but the differential expression seems to be estimated without regards to cancer type, which would certainly be highly biased and misleading. Why not do a simple regression of gene expression by ecDNA status, cancer type and select the genes that show significant coefficient for ecDNA status?*

Response. As requested by the reviewer, and in the more detailed questions below, we added an alternative model with a generalized linear model (GLM) analysis that controlled for tumor subtype. The method itself is described in the Methods section and pasted below. The GLM genes were tested along with the LFC, CorEx, Core genes as described in response to the previous question, and those results are now presented in Figure 2b and on pages 6 and 7 of the revised manuscript.

"We tested each of 16,309 genes independently in a separate logistic regression model using the glm() function in the R stats package (v4.2.0), and retained genes that were significant (p-value 0.01). Specifically, the model was defined as glm($y \sim gj + tt$, data = M, family = binomial(link = 'logit')), where y is the response vector where yi=1 if sample $i \in \{1, \ldots, 870\}$ is ecDNA(+) and yi =0 otherwise, gj is the vector of expression values for gene $j \in \{1, \ldots, 16309\}$ in samples $i \in \{1, \ldots, 870\}$, t is the covariate vector representing the tumor subtypes of samples $i \in \{1, \ldots, 870\}$, and M is the data matrix containing values of gene expression, tumor subtype, and ecDNA status for all samples. The equation for the binomial logistic regression

described above pp is formulated as $log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$, where p is the probability

that the dependent variable y is 1, X are the independent variables, and β are the coefficients of the model. In this case, k=1 represents independent variable gene j and k=2 represents the tumor subtype covariate t. Of the 16,309 genes tested independently, 3,012 genes were significant at pvalue<0.01."

> *(C) After identifying key features (which the authors inappropriate imply to be causal) they perform a series of enrichment/correlative analysis.*

Response. We have reviewed the document to ensure that we did not use the word 'causal.' If the reviewer can point to specific text, we are happy to change the phrasing.

> *(a) It is known that ecDNA status associates with poor survival, and so are cell cycle related signal. Then the association between Boruta genes and those processes is entirely expected. Is it not? The same goes for downregulation of immune processes.*

Response. We agree with the reviewer that cell cycle related signals and immune related signals are associated with low survival, and so does ecDNA. However, many cellular processes could be associated with low survival (including for example, metabolic processes, protein and DNA biosynthesis, etc.). The unexpected part is that there appear to be only 4 major processes that are upregulated in ecDNA(+) cancers relative to ecDNA(-) cancers, and only one (immune response) that is downregulated.

> *(b) The association with DSB specifically is interesting. Further analysis or discussion of why this should be would strengthen the work.*

Response. We thank the reviewer for their comment, and agree with their perspective. Note that we devoted a fair amount of text to analysis of DSB pathways. Specifically, we parsed the 4 main pathways in Figure 3b, and found our data to suggest that many genes in the classical nonhomologous end joining repair pathway are down-regulated in ecDNA(+) samples relative to ecDNA(-) samples. In contrast, Alternative end-joining and homology directed repair pathways are upregulated. This is a surprising result because c-NHEJ is considered to be an important mechanism of DSB repair. We have some lines in the discussion that address this:

"The DNA damage genes are broadly up-regulated in ecDNA(+) samples, especially in double-strand break repair. Within this broad category of mechanisms, our analysis suggests that alternative DSB repair pathways such as Alt-EJ are preferred relative to classical NHEJ. This is consistent with previous observations of small microhomologies at breakpoint junctions, and has important implications in therapeutic selection that will need to be validated in future experimental studies. We note, however, the microhomology analyses typically study breakpoint junctions, and might ignore double-strand breaks in non-junctional sequences which could be observed, for example at replication-transcription junctions."

We note that additional experimental work to corroborate these findings is significant effort and will be part of ongoing research in our collaborators' laboratories.

> *(c) On page 15, second paragraph, when providing the up versus down CorEx genes, please also provide up versus down for non-CorEx genes as well to get a sense of magnitude.*

Response. We thank the reviewer for the comment. We note that Supplementary Table S15 has the complete contingency tables as well as the Fisher Exact Test statistic for all categories. For the specific categories mentioned in the paper, the chi-square tables are reproduced below. As we are citing TableS15 (containing all numbers and the statistic p-value) in the main text, we thought it was better to leave the text as it was.

Category: Inflammation (p-value: 0.005)

CorEx: 18 (UP), 76 (DOWN)

Non-CorEx: 325 (UP), 657 (DOWN)

Category: Leukocyte migration and chemotaxis (p-value: 0.03)

CorEx: 13 (UP), 49 (DOWN)

Non-CorEx: 213 (UP), 410 (DOWN)

Category: Lymphocyte activation (p-value: 0.0075)

CorEx: 23 (UP), 75 (DOWN)

Non-CorEx: 334 (UP), 560 (DOWN)

Category: Cytokine production (p-value: 0.117)

CorEx: 6 (UP), 28 (DOWN)

Non-CorEx: 93 (UP), 208 (DOWN)

> *(d) The finding that Boruta genes are associated with high mutation burden is intriguing because in general mutation burden is associated with better survival and immunotherapy response. This counter-intuitive result should be scrutinized more to strengthen the work.*

Response. We agree with the reviewer that it is an intriguing observation. However, we are cautious in our interpretation. This is for the following reasons (all mentioned in the text):

(1) The total mutation burden was significantly higher in ecDNA(+) samples relative to ecDNA(-) samples (Fig. 5a). However, when controlling for cancer type, only glioblastoma, low-grade gliomas, and uterine corpus endometrial carcinoma continued to show differential total mutational burden (Fig. S7b).

(2) We tested if specific genes were differentially mutated between the two classes (Fig. 5b). For deleterious/high-impact mutations, TP53 was the only gene whose mutational patterns were significantly higher in ecDNA(+) compared to ecDNA(-) (OR 2.67, Bonferroni adjusted p-value 4.22e-07). BRAF mutations, however, were more common in ecDNA(-) samples and were significant to an adjusted p-value < 0.1 (OR 0.27).

(3) In response to another reviewer's comment, we also tested correlation with variant allele frequencies, and did not find any significant correlation except for TP53. We decided not to include that result in the paper.

These tissue specific cases might be confounding the main observation, but we have placed all of them together so that the reader can gain a better understanding. It is worth noting that the correlation between high TMB and immunotherapy response is also now controversial, and perhaps not true for all cancer types. See for example (https://www.annalsofoncology.org/article/S0923-7534(21)00123-X/fulltext), which suggests that this relationship is not true for Glioma, and in Glioma (which is ecDNA enriched), higher TMB is associated with worse immunotherapy response. Our results are consistent with that finding. We have modified the discussion paragraph to better reflect this.

"Mutation data alone does not provide as clear a picture of the genes involved in ecDNA maintenance. We did observe that the total mutation burden (TMB) was higher in ecDNA(+) samples. However, that relationship is much less clear after controlling for cancer type. High TMB has been positively correlated with sensitivity to immunotherapy52, and better patient outcomes; however, the gene expression patterns suggest that immunomodulatory genes are downregulated in ecDNA(+) samples, and patients with ecDNA(+) tumors have worse outcomes2. Notably, other results have suggested that the correlation between TMB and

response to immunotherapy is not uniform, and it can vary across different tumor subtypes53. Specifically, our data is consistent with previous results which showed that Gliomas with high TMB have worse response to immunotherapy relative to gliomas with low TMB53. In general, no collection of gene mutations was predictive of ecDNA status, although mutations in TP53 were more likely in ecDNA(+) samples, and perhaps are an important driver for ecDNA formation5."

> *(e) On page 17 "12 of the 47 genes not specifically enriching any known GO biological Process" is confusing. How can individual gene enrich for a GO process?*

Response. We agree that the statement was incorrectly phrased. We have changed it to state that "Only 12 of the 47 genes were not included in the gene sets of any enriched GO term."

> ***Reviewer #2 (Public Review):***
>
> *In their manuscript entitled "Transcriptional immune suppression and upregulation of double stranded DNA damage and repair repertoires in ecDNA-containing tumors" Lin et al. describe an important study on the transcriptional programs associated with the presence of extrachromosomal DNA in a cohort of 870 cancers of different origin. The authors find that compared to cancers lacking such amplifications, ecDNA+ cancers express higher levels of DNA damage repair-associated genes, but lower levels of immune-related gene programs.*
>
> *This work is very timely and its findings have the potential to be very impactful, as the transcriptional context differences between ecDNA+ and ecDNA- cancers are currently largely unknown. The observation that immune programs are downregulated in ecDNA+ cancers may initiate new preclinical and translational studies that impact the way ecDNA+ cancers are treated in the future. Thus, this study has important theoretical implications that have the potential to substantially advance our understanding of ecDNA+ cancers.*
>
> *Strengths*
>
> *The authors provide compelling evidence for their conclusions based on large patient datasets. The methods they used and analyses are rigorous.*
>
> *Weaknesses*
>
> *The biological interpretation of the data remains observational. The direct implication of these genes in ecDNA(+) tumors is not tested experimentally.*

Response. We agree with the reviewer that experimental tests would be ideal. Towards that, there are some challenges. The immune system genes cannot be tested in cell line models as they need a tumor microenvironment. Tests of DSB repair mechanisms and cell cycle control can be performed in cell-lines, but not with the TCGA samples which are not available. Some of our collaborators are actively working on these topics, but that extensive experimental work is beyond the scope of this paper.

> ***Reviewer #3 (Public Review):***
>
> *Summary:*
>
> *Using a combination of approaches, including automated feature selection and hierarchical clustering, the author identified a set of genes persistently associated with extrachromosomal DNA (ecDNA) presence across cancer types. The authors further validated the gene set identified using gene ontology enrichment analysis and identified that upregulated genes in extrachromosomal DNA-containing tumors are enriched in*

biological processes like DNA damage and cell proliferation, whereas downregulated genes are enriched in immune response processes.

Major comments:

(1) The authors presented a solid comparative analysis of ecDNA-containing and ecDNA-free tumors. An established automated feature selection approach, Boruta, was used to select differentially expressed genes (DEG) in ecDNA(+) and ecDNA(-) TCGA tumor samples, and the iterative selection process and two-tier multiple hypothesis testing ensured the selection of reliable DEGs. The author showed that the DEG selected using Boruta has stronger predictive power than genes with top log-fold changes.

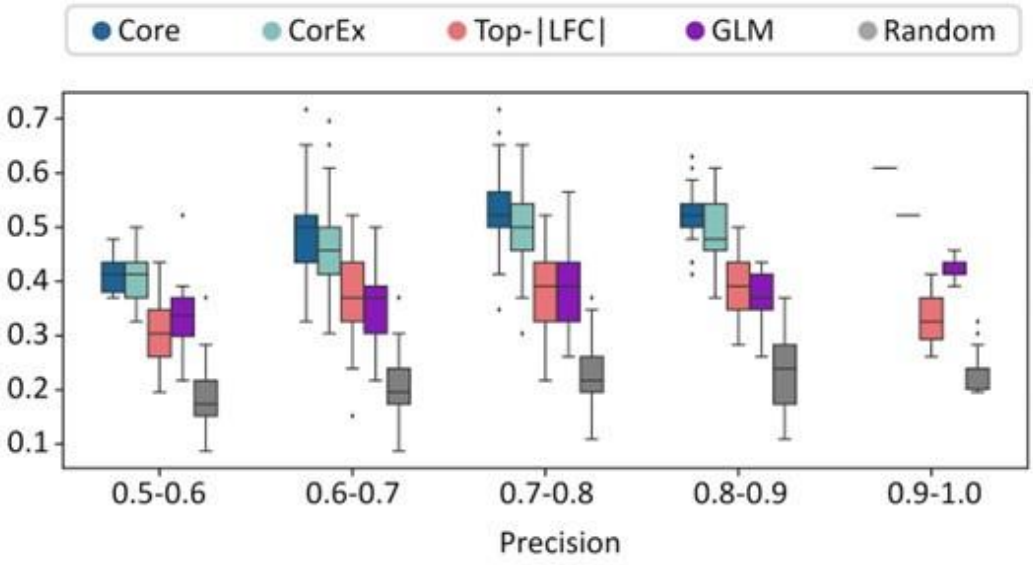
(2) The author performed a thorough interpretation of the findings with GO enrichment analysis of biological processes enriched in the identified DEG set, and presented interesting findings, including the enrichment in DNA damage process among the genes upregulated in ecDNA(+) tumors.

(3) Overall, the authors achieved their aims with solid data mining and analysis approaches applied to public data tumor data sets.

(4) While it may not be the scope of this study, it will be interesting to at least have some justification for choosing Boruta over other feature selection methods, such as Recursive Feature Elimination (RFE) and backward stepwise selection.

Response. We actually agree with the reviewer that some other feature selection methods could work just as well, and note that the Boruta analysis is not our creation, but a published feature selection method (Kursa, Journal of Statistical Software September 2010, Volume 36, Issue 11). We use Boruta to identify relevant genes, but the bulk of the paper is to understand the biological processes driven by that gene selection. Even if we had chosen another method that performed slightly better, it likely would not change the main conclusions. However, to address the reviewers concerns on over-reliance on one method, we added a different gene list created by a generalized linear model analysis, with the goal of checking if the expression of a gene could predict the ecDNA status of the sample after controlling for tumor subtype. Thus, we tested 5 different genelists in terms of their power in predicting ecDNA. While none of the lists is a great predictor of ecDNA status, the Core and CorEx gene lists are significantly better than the other lists. The Figure below replaces the previous Figure panels 2b and 2c.

Author response image 4.



> *(1) The authors showed that DESEQ-selected DEGs with top log-fold changes have less strong predictive power and speculated that this may be due to the fact that genes with top log-fold changes (LFC) are confined only to a small subset of samples. It will be interesting to select DEGs with top log-fold changes after first partitioning the tumor samples. For example, randomly partition the tumor samples, identify the DEGs with top LFC, combine the DEGs identified from each partition, then evaluate the predictive power of these DEGs against the Boruta-selected DEGs.*

Response. This is a great comment. We added a generalized linear model test for selecting genes whose expression is predictive of ecDNA status. The GLM list described above uses a standard methodology (Analysis of Variance) controls for tumor type as a covariate, and its predictive performance is only slightly better than the Top-|LFC| genes, while improving over a random gene set.

> *(2) While the authors showed that the presence of mutations was not able to classify ecDNA(+) and (-) tumor samples, it will be interesting to see if variant allele frequencies of the genes containing these mutations have predictive power.*

Response. This is a great suggestion. To address the reviewer's question, we used allelic counts (REFs and ALTs) information from the MC3 variant callset, and calculated allele frequencies of all variants from samples where ecDNA status was available. Next, we conducted a Wilcoxon rank-sum test between VAFs of the ecDNA(+) group and VAFs of the ecDNA(-) group for every mutated gene. We found 1,073 genes with $p<0.05$, but among them, only TP53 passed the multiple testing correction (padj<0.05, Benjamini-Hochberg). As the results are identical to the tests based solely on presence of mutations, we decided not to include this data.

> ***Reviewer #1 (Recommendations For The Authors):***
>
> *(A) The presentation should be substantially streamlined.*

> *(B) Preferably use a more intuitive simpler ML approach with fewer parameters to make it more credible. Because there are relatively few samples across numerous cancer types with greater variability in representation, a simpler procedure with transparent controls will be more convincing.*

Response. We accept the reviewer's criticism in that other statistical or ML approaches could potentially have been used, and that some are simpler. However, the test used here directly addresses the question: Find a collection of genes whose expression value is predictive of ecDNA status in the sample. Because the underlying method in the Boruta analysis uses random forests, it can test predictive power without relying on a linearity assumption implicit in other methods. In this revision, we also compare against a Generalized Linear Model (regression analysis) and show that it is less suited to the specific task above. We address the reviewer concerns about specific parameter choices by showing robustness to the specific parameter. All details are provided in the initial questions, and in the revised manuscript.

> *(C) Avoid using any term implying causality unless you can bring in direct experimental evidence (e.g. mutagenesis experiment followed by ecDNA measurement. Some places you use the word 'maintain ecDNA' and other places 'ecDNA impact'. But these are all associations. How can you distinguish causal genes from downstream effects without additional data?*

Response. We note that the word causal does not appear anywhere in the manuscript, and was not intended. Additionally we have revised the manuscript and are open to specific changes requested by the reviewer or the editors.

> *(D) Along these lines, if Boruta genes are indeed causal, one would expect Boruta-Up genes to be amplified more than expected in the ecDNA+; converse for Boruta-down genes.*

Response. We did not understand the reviewer's question. By "amplified," if the reviewer means "amplification of transcript level," then that is exactly what the Boruta analysis is showing. Specifically, for each gene, we have the ability to pick a transcript level cut-off 't' so that samples in which the expression is higher than t are more likely to be ecDNA(+). However, we are not claiming that there is causality, just that the transcript level is (weakly) predictive of the ecDNA status of the sample.

> *(E) A strawman control should be a simple regression-based gene identification that controls for ecDNA status and cancer type.*

Response. We agree that this was a very good suggestion. In the revision, we have applied a GLM, which controls for tumor type. Thus, we have 5 gene-lists (including the Core and CorEx genes). As described in the revised manuscript but also in response to the main comments above, none of the lists are a great predictor. However, the CorEx and Core genes are significantly better at predicting ecDNA status of a sample.

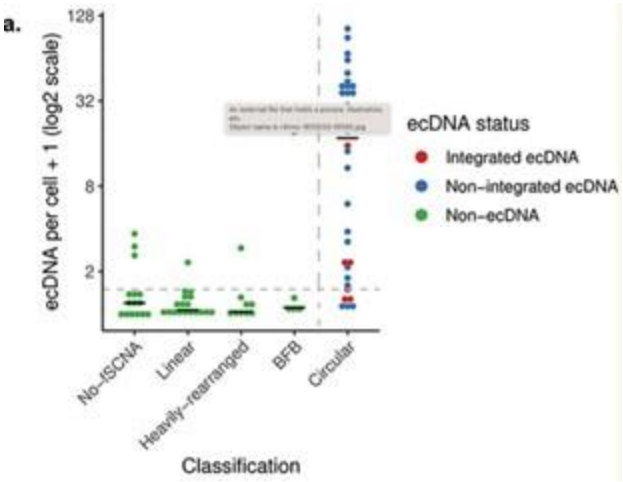
> ***Reviewer #2 (Recommendations For The Authors):***
>
> *Comments*
>
> *(1) The analysis hinges on a classification of tumors into ecDNA(+) and ecDNA(-) using AmpliconClassifier. It would be good to know how robust the outcomes are with respect to the performance of AmpliconClassifier - how many false positives and negatives will*

> *AmpliconClassifier generate on this dataset and how would this influence the CorEx genes?*

Response. This is a very reasonable request. AA has been extensively tested on established cell-lines for its ability in predicting ecDNA status, and this information is published in multiple venues, including Kim, Nature genetics 2020, and shows precision 85% for recall 83%. For completeness, we have reproduced the relevant plot from that paper here, and the relevant text here, but are not including it in the manuscript.

"To evaluate the accuracy of the AmpliconArchitect predictions, we analyzed whole-genome sequencing data from a panel of 44 cancer cell lines, and examined tumor cells in metaphase. We used 35 unique fluorescence in-situ hybridization (FISH) probes in combination with matched centromeric probes (81 distinct "cell-line, probe" combinations) to determine the intranuclear location of amplicons (Supplementary Table 2). Following automated analysis >1,600 images, we observed that 85% of amplicons characterized as 'Circular' by whole genome sequencing profile demonstrated an extrachromosomal fluorescent signal, representing the positive predictive value. Of the amplicons corresponding to extrachromosomally located FISH probes, 83% were classified as Circular, representing the sensitivity (Extended Data Fig. 1A)."

Author response image 5.

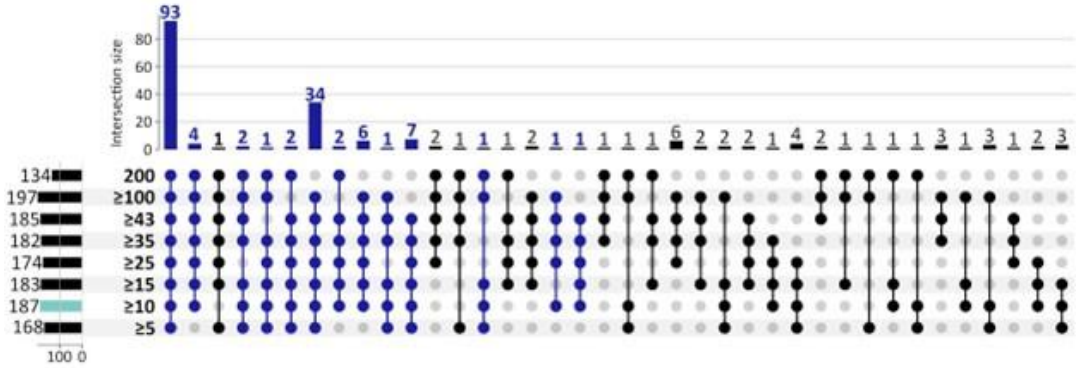


> *(2) It is unclear why genes are labeled Boruta genes when they are present in 10 out of 200 runs, this seems like an unexpectedly low number. How did the authors arrive at this number? Do the authors have any ground truth to estimate how well Boruta works in this setting and implementation?*

Response. This is a great question and asked by another reviewer as well. Given the weakness of an individual gene as a classifier, its repeated selection in multiple Boruta trials is already a significant event. By requiring a gene to be picked in 5% of the trials (10/200), we were selecting a small, but more robust list of genes. However, to further explore the reviewer's concerns, we also applied 8 other selection criteria ranging from 5 (of 200 Boruta trials) to 200 of 200 Boruta trials. See Figure below. The number of CorEx genes expectedly decreases with increasing stringency. However, of the 187 GO terms that were enriched by UP-genes, 93 terms (50%) were enriched regardless of the cut-off (see Figure below), and 153 terms (82%) were enriched in at least 5 of the 8 cut-offs. Given that the remaining analysis works on the hierarchy of GO terms and finds 4 GO-categories (Mitotic Cell Cycle, G1/S, G2/M; cell-division;

DSB DNA Damage response; and the HOX Gene cluster) enriched by UP-regulated genes, those conclusions would hold regardless of the specific cut-off.

Author response image 6.



The number of GO terms that were enriched by DOWN-regulated genes is smaller, only 73, and falls rapidly for higher cut-offs, with 25 at a cut-off of 15. Therefore we see fewer terms enriched for more stringent cut-offs. However, they all support immune processes. These results do suggest that there are fewer genes that are consistently down-regulated in ecDNA(+) cancers, and expression change in a small number of genes may be sufficient to promote conditions for ecDNA.

We have added the figure as a supplemental figure and have added the following text to the manuscript on pages 17 and 18.

“Any CorEx gene is either a Core gene that was selected as a feature in at least 5% of 200 Boruta trials, or be highly co-expressed with a Core gene. Because the selection criterion of 5% is arbitrary, we also tested robustness with 8 other cut-offs ranging from 5-of-200 to 200-of-200 Boruta trials. The number of CorEx genes expectedly decreases with more stringent cut-offs.

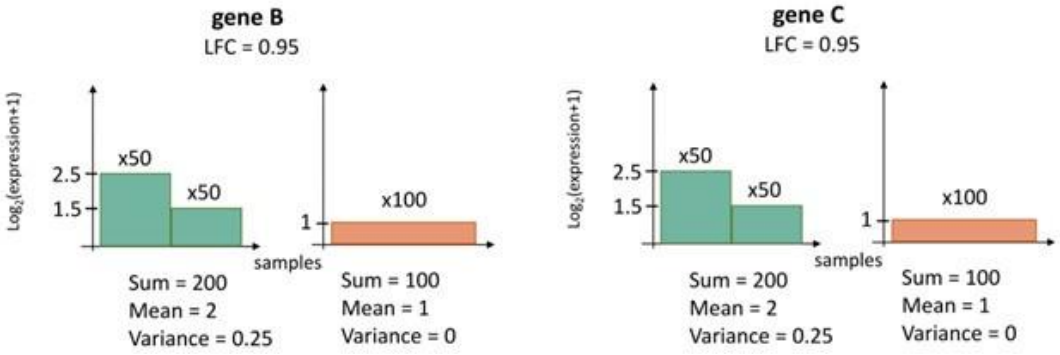
However, of the 187 GO terms that were enriched by 262 CorEx UP-genes using 10 of 200 Boruta trials as the selection criteria, 93 terms (49.7%) were enriched for each cut-off (Fig. S9), and 155 terms (82.9%) were enriched in at least 5 of the 8 cut-offs. Given that our subsequent analyses utilized the hierarchy of GO terms and identified 4 GO-categories enriched by UP-regulated genes, the conclusions would hold regardless of the specific cut-off.”

> *(3) Authors extend the core gene set with co-expressed genes, arguing that "gene C" would not add predictive power in addition to "gene B" and is therefore not identified as a Boruta gene. However, from its description in the manuscript (summarized: "Boruta [...] selects the highest feature importance score, s, of shadow features as a cut off, and returns features with a higher score than s."), it isn't immediately obvious to me why Boruta would not return both genes B and C. Maybe the authors could explain this better.*

Response. We consider the following.

(1) Consider 100 ecDNA(+) and 100 ecDNA(-) samples. Let the expression levels of genes B and C in the data-sets be as described in the figure below; y-axis is the gene expression, and x-axis is just a listing of all samples, with green color denoting ecDNA(+) samples and orange color denoting ecDNA(-) samples.

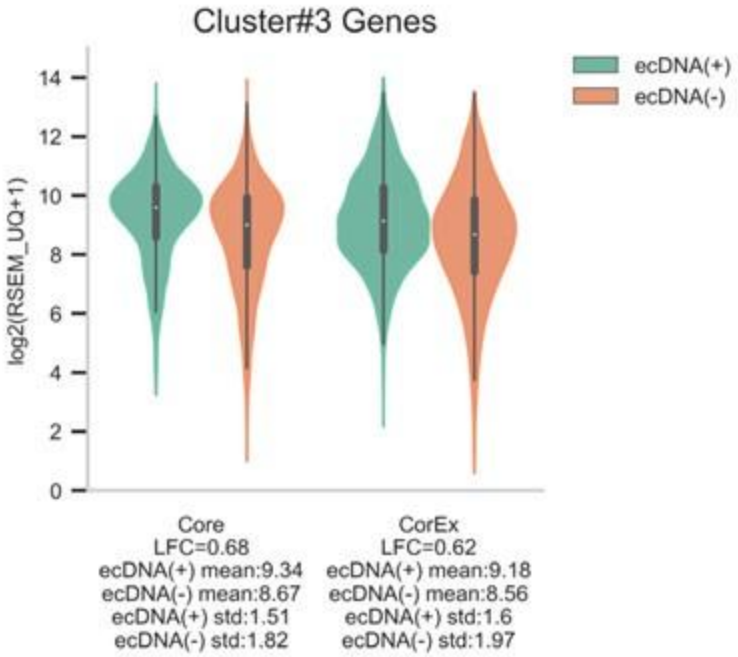
Author response image 7.



(2) Then, if we choose gene B and a transcript level of 1.25, we have a perfect prediction of ecDNA status because all samples where gene B has a transcript level higher than 1.25 are ecDNA(+) and otherwise they are ecDNA(-). Similarly, using Gene C, we can get perfect predictions. Thus, when Boruta has to select a gene, it will pick either Gene B or Gene C, because picking both will not improve prediction. We can therefore use Boruta to pick one gene, and then co-expression clustering to pick the other gene.

As an example, cluster #3 consists of 21 genes that were up-regulated in ecDNA(+) samples and enriched in cell-cycle related biological processes (Table S3). While these genes were expressed similarly in ecDNA(+) samples, and separately, in ecDNA(-) samples, out of the 21 genes, only 9 genes were selected in at least 10 out of 200 Boruta trials (i.e., Core genes). Of the 12 remaining genes (i.e., CorEx genes), 8 genes were not selected by the Boruta method at all, 3 genes were selected in less than 5 out of 200 Boruta trials, and 1 gene was selected in 9 out of 200 Boruta trials.

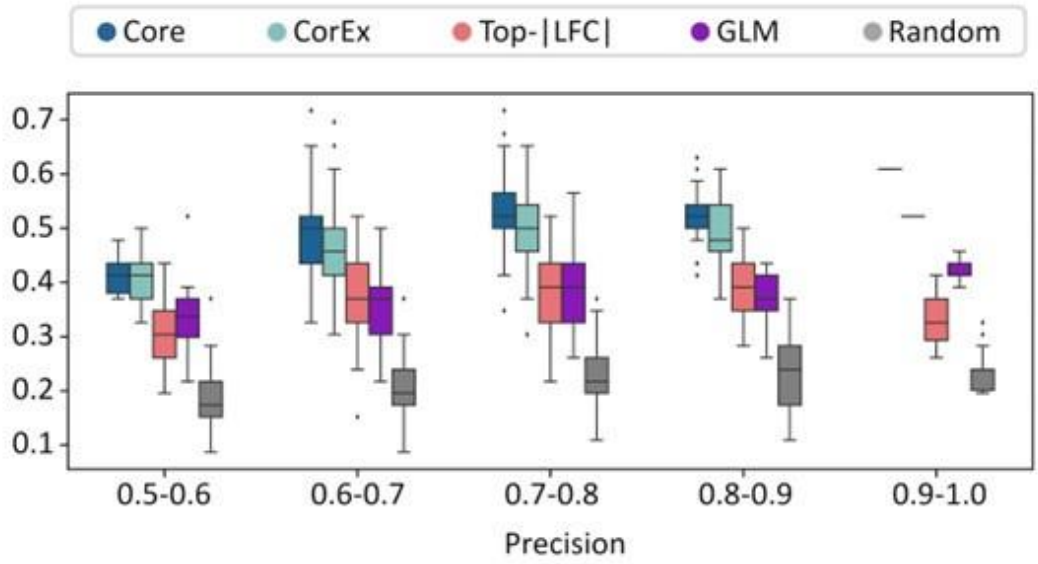
Author response image 8.



> *(4) In Fig 2a, I would like to see the variability of the precision and recall in the main text, not only the maximum values. Authors could plot mean + standard deviation for precision and recall separately, or use S2a/b.*

Response. We have replaced Figures 2b and 2c with a combined figure (Fig. 2b) that gives a box-plot describing the distribution of recall values for 5 gene lists: four from the original manuscript, and another gene list created using a Generalized Linear Model (GLM).

Author response image 9.



> *(5) Since the authors analyze bulk RNA, the gene expression signatures they notice could, in principle, originate from non-tumor cells as well. I do not believe this is the case, however, the paper would be strengthened by an analysis that shows that the difference in expression patterns of the Corex genes between ecDNA(+) and ecDNA(-)-samples does come from tumor cells. One way of showing this would be by using single-cell mRNA-sequencing data, and another way of showing this would be to show that Corex gene-expression correlates with tumor purity in bulk samples.*

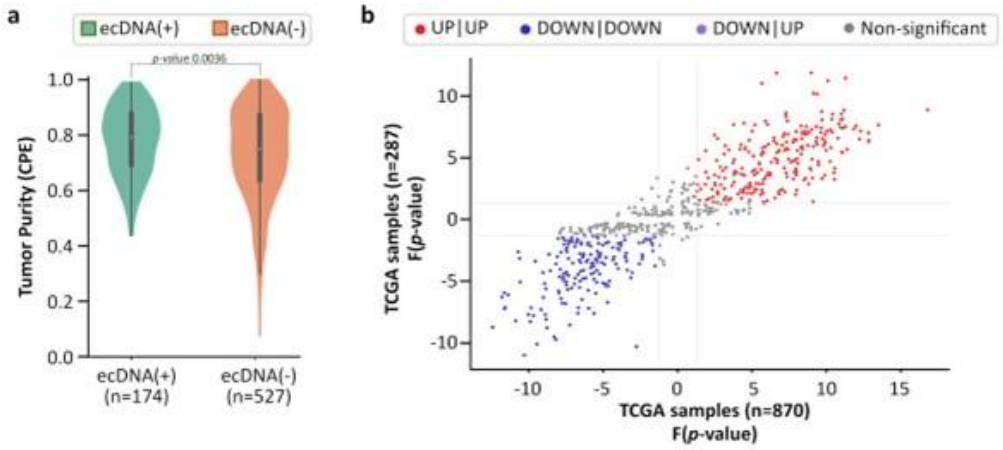
Response. The reviewer is correct. Unfortunately, our analysis requires data with whole-genome sequencing (WGS) for ecDNA prediction, as well as RNA-seq for transcriptome profiling. The TCGA data-set is the only available data-set with a significant number of samples that includes both WGS and RNA-seq. They have not made tissue samples available for scRNA analysis, to our knowledge. The reviewer raises an important question regarding purity, but testing if CorEx gene expression correlates with tumor purity would require a large range of purity values, something that scientists would avoid when collecting samples.

However, the presence of non-cancer tissue (impurity) could reduce sensitivity of ecDNA detection, and therefore, change the results. To better investigate this, we started with a publication that investigated multiple tumor purity metrics and devised a composite score (CPE; Aran et al., 2015). Using their composite tumor purity, we find that ecDNA(-) samples have slightly lower purity than ecDNA(+) samples (p-value 0.0036; Fig. S2a).

This result is not surprising because one would expect lower detection of ecDNA in less pure samples. The presence of undetected ecDNA in ecDNA(-) samples would confound the results

by reducing the discriminating power of genes, but would not give false results. To test this, we measured the expression directionality in CorEx genes in all samples versus samples which had a high tumor purity (CPE 0.8). The results suggest that the p-values of directionality in the pure samples were highly correlated with the expression data from all samples (Fig. S2b).

Author response image 10.

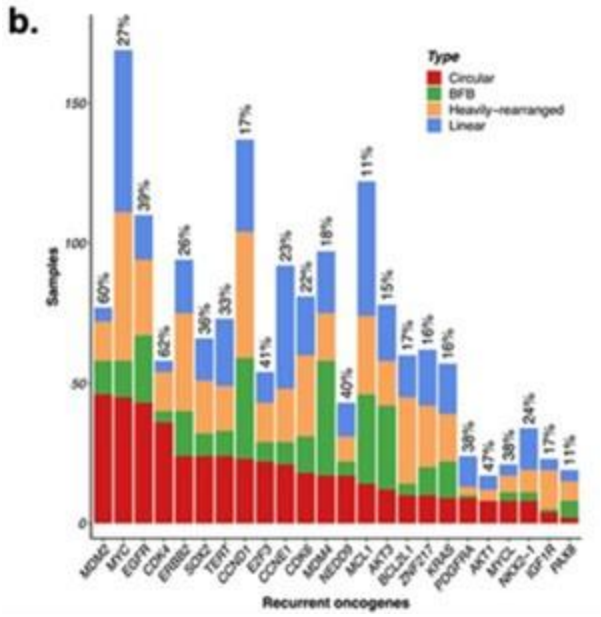


> *(6) The biological interpretation of the data remains a bit too observational. Can the authors offer an interpretation of the enriched GO terms? And are any of these genes already implicated in ecDNA(+) tumors?*

Response. To answer the second question first, prior to our study, the focus was on genes that were amplified on ecDNA. Indeed many oncogenes known to be amplified in cancer are in fact amplified on ecDNA (Turner, Nature 2017, Kim Nature genetics 2020). This study is unique in that it identifies genes whose expression values are predictive of ecDNA(+) status. The Figure below lists 24 genes most frequently amplified on ecDNA from Kim, Nature Genetics 2020. With the exception of EGFR and CDK4, none of these 24 genes was included in the list of the 65 genes reported by us as the most frequently selected genes in the Boruta trials (lowest harmonic rank). Thus, most persistent CorEx genes do not lie on ecDNA. However, they all play important roles in biological processes relevant to cancer pathology including Immune Response, Mitotic cell Cycle, Cell division, and DSB repair. We agree with the reviewer that the results are observational (although statistically significant in populations), and some of our collaborators are actively working to experimentally validate some of these genes. The experimental work, however, is beyond the scope of this paper.

We have added the following statement to the manuscript. “Notably, of the 24 genes most frequently expressed on ecDNA,2 only EGFR and CDK4 were included in the list of 65 genes, suggesting that the most persistent CorEx genes do not themselves appear frequently on ecDNA.”

Author response image 11.



> ***Reviewer #3 (Recommendations For The Authors):***
>
> *Minor comments:*
>
> *(1) The authors performed gene ontology enrichment test but referred to it as gene set enrichment analysis. Usually gene set enrichment analysis does not refer to Fischer's exact test-based analysis but rather the one described in Subramanian et al 2005. The term correction should be made to avoid confusion.*

Response. We have rephrased text in the manuscript to prevent confusion between enrichment analysis on gene sets using an one-sided Fisher's exact test and the Gene Set Enrichment Analysis (GSEA) method that exists as a software. We have also revised the header in the methods section from "Gene set enrichment analysis" to "Gene Ontology (GO) enrichment analysis".

> *(2) A couple of figures could use more detailed labels and captions. In Figure 2c, it is unclear what the numbers 100 and 54 right next to the Cliff's Delta heatmap indicate. In Figures 3a and 4a, it is not immediately clear what the barplot on top of the heatmap indicates and there is no label for the y-axis.*

Response. These are good suggestions, and we have added descriptions to the figure captions.