Exploring the druggable space around the Fanconi anemia pathway using machine learning and mechanistic models

Marina Esteban-Medina, María Peña-Chilet, Carlos Loucera and Joaquín Dopazo

Abstract

Background: In spite of the abundance of genomic data, predictive models that describe phenotypes as a function of gene expression or mutations are difficult to obtain because they are affected by the curse of dimensionality, given the disbalance between samples and candidate genes. And this is especially dramatic in scenarios in which the availability of samples is difficult, such as the case of rare diseases.

Results: The application of multi-output regression machine learning methodologies to predict the potential effect of external proteins over the signaling circuits that trigger Fanconi anemia related cell functionalities, inferred with a mechanistic model, allowed us to detect over 20 potential therapeutic targets.

Conclusions: The use of artificial intelligence methods for the prediction of potentially causal relationships between proteins of interest and cell activities related with disease-related phenotypes opens promising avenues for the systematic search of new targets in rare diseases.

Keywords: Genomics, Big data, Machine learning, Fanconi anemia, Signaling pathways, Mathematical models

Background

With the extraordinarily fast increase in throughput that sequencing technologies underwent in the last years [1, 2], genomics has become a de facto Big Data discipline. Recent prospective studies have compared genomic data generation with other major data generators such as astronomy, twitter and youtube and have concluded that genomics is either on par with or, possibly even most demanding than the Big Data domains analyzed in terms of data acquisition, storage, distribution, and analysis of data [3]. Therefore, this seems to be the ideal scenario for the application of machine learning techniques, that have recently been successfully applied to many domains of medicine [4] such as radiology [5], pathology [6], ophthalmology [7], cardiology [8], etc. However, in the case of human genomic data, most of the applications have been unsupervised class discovery approaches, using gene expression data for visualization, clustering, and other tasks, mainly in single-cell [9, 10] or cancer [11, 12], being supervised applications restricted to a few examples of relatively simple problems, in which a good balance between variables to predict and data available is satisfactory, such as inferring the expression of genes based on a representative subset of them [13] or predicting the activity status of Ras pathway in cancer [14]. Consequently, in spite of the wealth of genomic data available there is a lack of translational applications due to the fact that the most interesting predictive scenarios face a serious problem of potential overfitting. Thus, attempts to describe complex, multivariant phenotypes as a function of an undefined number of genes are hampered by the high number of variables (in the range of 20,000 genes [15]), which challenge many conventional machine learning (ML) approaches. Therefore, new strategies that exploit the enormous potential of ML applied to genomic Big Data in order to model diseases and discover new therapies are necessary.

An especially interesting use of genomic data is related with the application of ML to model the function of the
cell [16]. Such models form a natural bridge from variations in genotype (at the scale of gene activities) to variations in phenotype (at the scale of cells and organisms) [17, 18]. Despite, these models are based on yeast, an organism far simpler than human, and use yeast genomic data, which are far more abundant than human genomic data, the framework proposed is interesting not only because of the use of a causal link between genotype and phenotype but also because it is attained with a dimensionality reduction. Thus, mechanistic models of human cell signaling [19] or cell metabolism [20] can provide the functional link between the gene-level data available (gene expression) and the cell phenotype level, allowing the selection of specific disease-related cellular mechanisms of interest. In fact, mechanistic models have helped to understand the disease mechanisms behind different cancers [21–24], the mechanisms of action of drugs [19], and other biologically interesting scenarios such as the molecular mechanisms that explain how stress-induced activation of brown adipose tissue prevents obesity [25] or the molecular mechanisms of death and the post-mortem ischemia of a tissue [26].

Here we plan to use a mechanistic model of the molecular mechanism of a disease, Fanconi anemia (FA) (ORPHA: 84), a rare condition that causes genomic instability and a range of clinical features that include developmental abnormalities in major organ systems, early-onset bone marrow failure, and a high predisposition to cancer [27]. Signaling is known to play a relevant role in the disease and also defines its most characteristic hallmark: failure of DNA repair [28, 29]. In addition, it has been described that FA influences survival and self-replication of hematopoietic cells [30].

Solid tumors, usually with poor prognosis as tumor resection is the only therapeutic option given that patients do not tolerate chemotherapy or radiation, constitute one of the most relevant hallmarks of FA. With improved hematopoietic stem cell transplant protocols, FA patient survival has increased, leading to a progressively increased number of solid malignancies in adult patients. Therapeutic research is currently focused in targeted therapies for solid tumors as well as in preventive options in the context of drug repurposing [31].

At present, a detailed map of FA signaling is available in the Kyoto Encyclopedia of Genes and Genomes, KEGG, (03460) that can be used to derive a mechanistic model that relate gene expression to the activity of signaling circuits within the FA pathway that trigger cell activities related to FA hallmarks. These models can be used to investigate other molecules that could affect the activity of such circuits and therefore, presumably, to FA hallmarks. Therefore, these molecules are potential therapeutic targets. Since we are dealing with a rare disease, which typically are not considered as attractive business niches by pharmaceutical companies [32], we will restrict the search space to proteins that are already targets of approved drugs. Actually, here we are aiming for drug repurposing, that is, the discovery of new indications for drugs already used in the treatment of other diseases [33], an ideal strategy for rare diseases that accelerates enormously the evaluation of candidate molecules and simultaneously reduces failure risks [34].

The attainment of the relationships between candidate proteins for a new indication and the FA hallmarks poses a challenge that can be addressed with the appropriate ML method.

**Results**

**General approach**

Here we take advantage of the biological knowledge available on FA, as represented in the FA pathway. The FA pathway describes the functional interaction among genes that finally trigger, from six different circuits, cell functionalities related with DNA repair (see Fig. 1), a known FA hallmark. Since the disease condition involves the malfunction of one or several of these DNA repair cell functionalities, we hypothesize here that other genes that have an influence on the status of these functionalities might be playing the role of upstream regulators and therefore their potential modulator capacity could eventually make of them suitable therapeutic targets. In order to find druggable genes that could be playing a significant modulator role over FA hallmarks we use known drug target (KDT) genes listed in DrugBank [35] (Additional File 1). These genes are used to predict the activity of the signaling circuits triggering the FA hallmarks.

Since the FA pathway available in KEGG seems incomplete we first build a curated expanded version of the FA pathway (see below). Then, we search for potential known drug targets that affect the functionality of the FA pathway. Figure 2 summarizes the procedure followed: for each sample of each tissue available for each individual (over 11,000), the activity of the genes in the pathway is used to estimate the activity of the circuits contained in the FA pathway using Hipathia [21]. Then, across the 11,000 samples, the ML procedure tries to infer the circuit activities from the expression levels of the KDT genes external to the pathway.

**Building a curated Fanconi Anemia disease map**

Here we use as starting point the KEGG FA pathway (hsa03460). However, among the 54 genes present in the pathway (see Additional File 2), three known FA genes (MAD2L2, RFWD3 and XRCC2) described in Orphanet (ORPHA:84) were missing, which suggests that the FA KEGG pathway probably does not constitute an updated version of the current knowledge on FA. Therefore, we have derived a manually curated expanded version of the FA map. To achieve so we have used the package pubmed.mineR [36] with all the possible pairs of FA genes.
Fig. 1 Fanconi anemia curated map, based in the KEGG FA pathway. There are two protein complexes: RPA, composed of RPA1, RPA2, RPA3 and RPA4, and Core, composed of FANCM, FANCG, FANCI, FANCA, FANCB, UBE2T, STRA13, FANCC, FAAP24, HES1, FANCE, FANCF, BLM, RMI1, RMI2 and TOP3A. At the end of the effector nodes, whose names are taken for the circuits, a description of the main functionalities triggered by the signaling circuits can be found.

Fig. 2 Schema of the procedure followed for the analysis.
searching for direct functional interactions. The results confirmed all the gene-gene interactions described in the KEGG pathway and expanded the connections to the three genes not present in the KEGG version as well as discovered 12 new interactions among FA genes (see Table 1). Figure 1 depicts the FA pathway expanded by manual curation.

Interestingly, in spite of the small number of samples in the comparison, the use of a mechanistic model, built in Hipathia [48] with the curated FA pathway, to analyze an experiment that compares gene expression in bone marrow cells between normal volunteers and FA patients [30] (GSE16334) rendered a significantly different activity in two circuits: REV3L (FDR-adj. $p$-value = $5.1 \times 10^{-4}$) and the RPA complex (FDR-adj. $p$-value = $4.5 \times 10^{-3}$), as well as the MLH1-PMS2, almost significant (see Table 2) that could not be detected when using the original KEGG FA pathway. Therefore, the curated pathway demonstrates a better detection of the expected differential behavior between normal and diseased bone marrow tissue than the original FA pathway, directly taken from KEGG. Figure 3 shows the distributions of the activities of different FA pathway signaling circuits in healthy and FA bone marrow cells in which more pronounced differences in circuit activity can be visualized for the above-mentioned circuits (REV3L, the RPA complex and MLH1-PMS2). Actually, Additional File 3 shows the same distribution obtained for the original FA KEGG pathway, where some incoherence can be observed, such as the absence of activity in four of the seven circuits. Figure 4 shows the activity in different normal tissues, taken from GTEx, which include blood, a tissue affected by the disease, two tissues with a high rate of cell replication (skin and gastrointestinal), where DNA reparation is expected to play a relevant role, and another tissue with low rate of cell replication (brain). Unfortunately, there are no expression data for bone marrow, the main tissue affected by the disease, in GTEx. DNA reparation circuits show a slightly different activity in brain when compared to the rest of tissues in the case of the three FA circuits.

### Table 1

| NODE 1 | NODE 2 | INTERACTION | Ref. |
|--------|--------|-------------|------|
| MAD2L2 | REV3L  | binding     | [37] |
| RFWD3  | RPA1   | binding/association | [38] |
| XRCC2  | RAD51C | activation   | [39] |
| REV1   | MAD2L2 | binding/association | [37] |
| FANC   | REV1   | activation   | [40] |
| POLK   | REV1   | binding/association | [41] |
| BRCA1  | REV1   | activation   | [42] |
| BRIP1  | BRCA1  | binding/association | [43] |
| PALB2  | BRCA2  | binding/association | [44] |
| PALB2  | BRCA1  | binding/association | [45] |
| FANCA  | BRCA1  | binding/association | [46] |
| FANCD2 | BRCA1  | binding/association | [47] |

### Exploring the druggable space of influence over the FA pathway

As sketched in Fig. 2, the ML strategy was applied to detect proteins whose activity was able of predicting the activity of the FA circuits that trigger the FA hallmarks. The initial search space was restricted to KDTs extracted from DrugBank (See Additional File 1). The cross-validation of the relevance values (Fig. 5) rendered a threshold of 0.006, above which the most relevant genes presented a stable value.

The importance of the genes selected by the ML strategy is strongly supported by a high predictive performance across all the splits, as can be seen in Fig. 6. The distribution of the $R^2$ score for each signaling circuit of the FA curated pathway across all the training/test splits have in all the cases a value close to 1 (note that the $R^2$ score goes from -infinite to 1, where 0 represents a model that always predicts the mean for each task and a perfect model has a score of 1).

A total of 17 genes resulted to have a relevance over the 0.006 threshold (See Table 3). Additional File 4 contain details on the drugs targeting these proteins.

### Discussion

Mechanistic models and machine learning approach used

Supervised ML applications in the case of human genomic data aiming to find genes potentially causal of phenotypes have restricted to a few cases in quite simple scenarios, such as the inference of very simple (and univariate) phenotypes, such as the activity status of Ras pathway in cancer [14]. Here we aimed to approach the pathologic phenotype problem in more detail, trying to capture the complexity of the molecular mechanism of the disease. To achieve so, we have used signaling circuit activities inferred by mechanistic models, as proxies of disease-related cell functionalities triggered by them. Such mechanistic models use gene expression data to produce an estimation of profiles of signaling or metabolic circuit activity within pathways [20, 24] and have been used to describe the molecular mechanisms behind different biological scenarios such as the explanation on how stress-induced activation of brown adipose tissue prevents obesity [25], the common molecular
mechanisms of three cancer-prone genodermatoses [49] or the molecular mechanisms of death and the post-mortem the ischemia of a tissue [26]. Moreover, recent benchmarking of mechanistic modeling methods shows how Hipathia clearly outperform to other competing method [50].

To assess the suitability of the expanded FA pathway, we have analyzed the distribution of the activity of its circuits once modeled in Hipathia. As expected, the overall activity in blood, skin and gastrointestinal tissues is higher than that of brain cells, due to its higher replication rate (Fig. 4). However, brain tissue also exhibits pathway activity to some extent, which can be explained by the involvement of FA pathway in DNA repair, since brain cells have high level of metabolic activity and use distinct oxidative damage repair mechanisms to remove DNA damage [51]. We also observed in Fig. 3 that RAD51C and REV3L circuit activities derived from the expanded FA pathway are, contrarily to the results obtained from KEGG FA pathway (Additional Fig. 3), significantly lower in FA patients than in healthy donors. This observation is coherent with the fact that these circuits are involved in DNA crosslinking repair during homologous recombination, a mechanism that has been demonstrated to be damaged in FA patients [39]. An interesting advantage of using mechanistic pathway models is that focusing on the pathways and circuits directly related to the disease hallmarks is straightforward. The analysis of the FA dataset [30] renders a high number of genes deregulated, which affect more pathways. However, many of the affected functionalities are consequences of the disease hallmarks or unrelated to them [52].

Therefore, the mechanistic models of the extended FA pathway offer the possibility of discovering what protein

Table 2 Differential circuit activity in a comparison of healthy versus FA bone marrow cells. Circuits are named after their effector nodes (see Fig. 1)

| CIRCUIT            | Activation | Statistic | p-value | FDR adj. p-value |
|--------------------|------------|-----------|---------|------------------|
| RAD51              | UP         | 0.615     | 0.538   | 0.659            |
| MLH1-PMS2          | UP         | 2.400     | 0.016   | 0.067            |
| REV3L              | DOWN       | -3.789    | 3.917 × 10⁻⁵ | 5.092 × 10⁻⁴ |
| RAD51C             | DOWN       | -1.924    | 0.056   | 0.162            |
| RPA*               | UP         | 3.412     | 6.923 × 10⁻⁴ | 4.500 × 10⁻³ |
| FANCM-STRA-FAAP24  | UP         | 1.885     | 0.062   | 0.162            |
Fig. 4 Observed distribution of circuit activities in blood, a tissue affected by the disease, two tissues with a high rate of cell replication (skin and gastrointestinal), where DNA reparation is expected to play a relevant role and another tissue with low rate of cell replication (brain).

Fig. 5 Distributions of the cross-validation of the relevance values for the top 50 most relevant genes ordered by their mean. Above the relevance value of 0.006 the relevance rendered by the ML procedure and the means obtained from the cross-validation are consistent. Then this value is taken as a threshold.
Fig. 6 shows the distribution of the $R^2$ score for each signaling circuit of the FA pathway across all the training/test splits. The $R^2$ score goes from $-\infty$ to 1, where 0 represents a model that always predicts the mean for each task and a perfect model has a score of 1.

Table 3: List of most relevant genes (relevance > 0.006) obtained by the model. Drug IDs in bold are approved for use according to DrugBank database.

| GENE NAME                                | SYMBOL | ENTREZ ID | RELEVANCE | TARGETING DRUGS (DrugBank ID)       |
|------------------------------------------|--------|-----------|-----------|-------------------------------------|
| NIMA related kinase 2                   | NEK2   | 4751      | 0.097324  | DB07180, DB12010                    |
| DNA topoisomerase II alpha               | TOP2A  | 7153      | 0.078623  |                                      |
| baculoviral IAP repeat containing 5      | BIRC5  | 332       | 0.052406  |                                      |
| centromere protein E                     | CENPE  | 1062      | 0.036961  |                                      |
| polo like kinase 1                       | PLK1   | 5347      | 0.036159  | DB06897, DB06963, DB07789           |
| cyclin dependent kinase 1                | CDK1   | 983       | 0.022697  | DB05037, DB06195                    |
| glutamate ionotropic receptor NMDA type subunit 1 | GRIN1 | 2902     | 0.019528  | DB01931, DB04620, DB05824, DB06741, DB09409, DB09481 |
| cholinergic receptor nicotinic beta 2 subunit | CHRNA2 | 1141     | 0.013228  | DB05855                             |
| synaptosome associated protein 2S        | SNAP2S | 6616      | 0.012799  |                                      |
| enhancer of zeste 2 polycomb repressive complex 2 subunit | EZH2 | 2146 | 0.012543 | DB12887, DB14581                    |
| methylenetetrahydrofolate dehydrogenase, cyclohydrolase and formyltetrahydrofolate synthetase 1 | MTHFD1 | 4522 | 0.012111 | DB00116, DB02358, DB04322          |
| thymidylate synthetase                   | TYMS   | 7298      | 0.009462  | DB00293, DB00322, DB00432, DB00440, DB00544, DB00642, DB01101, DB05116, DB05308, DB05457, DB07577, DB08478, DB08479, DB08734, DB09256 |
| serpin family E member 1                 | SERPINE1 | 5054   | 0.009206  | DB05254                             |
| cytochrome c oxidase subunit I           | COX1   | 4512      | 0.008027  | DB09140                              |
| retinoic acid receptor alpha             | RARA   | 5914      | 0.007607  |                                      |
| sodium voltage-gated channel alpha subunit 2 | SCN2A | 6326 | 0.006728 | DB13520                              |
| kinesin family member 11                 | KIF11  | 3832      | 0.006366  | DB03996, DB04331, DB04900, DB07064, DB08032, DB08033, DB08037, DB08198, DB08239, DB08244, DB08246, DB08250 |
activities potentially affect the different pathway activities that trigger FA hallmarks, which provide a mechanistic link between such proteins and the disease phenotype. However, finding these relationships constitutes a complex problem that involve multiple variables (here KDT proteins) to predict multiple outputs (here signaling circuit activities related with DNA repair, a FA hallmark) that can be formulated as multi-output regression problems (MOR), also called multi-task learning or vector valued regression. MOR is a fundamental problem in machine learning as it deals with the ability to predict multivariate responses with a single model, instead of learning one model per output, the classic single output regression (SOR) scenario, e.g. conventional univariate regression. The MOR scenario has several advantages over SOR: on the one hand, contrarily to the case of SOR, where each variable to predict is treated as independent (uncorrelated), in MOR, the variables to predict, the circuit activities, are correlated, which makes sense from a biological point of view. In other words, SOR, requires a different set of hyper-parameters (i.e. a different model) for each variable, leading to several training/testing/validation scenarios with different features learned, while in the MOR learning framework a unique model (only one set of hyper-parameters) is used to predict all the output variables (circuits) at once, with the ability to exploit and learn the shared patterns between them. Therefore, the MOR scenario provides an ideal framework to properly address hypothesis from a systems biology point of view given that it assumes that the response variables, here the different signaling circuits in the FA pathway are (or can be) interconnected. An additional advantage of using mechanistic models is that, by accurately defining the functional space of interest (the FA hallmarks described in the FA pathway), the number of circuits involved in their activity results relatively low, which constitutes a reduction of the dimensionality of the output space based on biological knowledge.

Here we used Random Forests (RF) [53], an ensemble of decision trees that aggregates the output of each estimator in order to stabilize and improve the prediction power. RFs and other tree-based ensembles have been proven to be extremely well suited for interpretable machine learning across different systems biology scenarios [54]. Tree-structured methods (TSM) provide a set of interpretable rules by splitting data into sample/target-wise homogenous groups and averaging the results. However, the predictive performance of a single decision tree is subpar when compared to other methods, such as Support Vector Machines, mostly due to the fact that a tree must make several sequential choices based on a subset of the data and one incorrect decision can impact the rest of the sequence, thus propagating the error. To improve the performance of a decision tree, several strategies have been proposed, the most notable among them are those based on building an ensemble of trees, where several trees (from hundreds to thousands) are fitted on different partitions of the training data or under different conditions, and then combined in order to achieve a better prediction capability [55]. On top of this, RFs are particularly well suited for the analysis of genomics datasets [56, 57] due to its robustness in scenarios affected by the curse of dimensionality.

Although one key advantage of RFs is its ability to produce good enough results with minimal hyperparameter search (given a sufficiently large number of trees are trained), in some circumstances the hyperparameter space must be properly optimized in order to obtain a good set of results [58]. Our problem setup is one of such cases, where a large number of highly correlated predictor variables (gene expression) interact with a multivariate response with many self-interactions (pathway circuit activities). To overcome such difficulties, we make use of Tree-structured Parzen Estimator (TPE) [59], a Sequential Model-based Global Optimization strategy for hyperparameter optimization. The base learners of a RF, the decision trees, can be easily extended to the multi-output scenario [60] by introducing a covariance weighting to the splitting criterion with the aim of finding a representation of homogeneous clusters with respect to both the predictor and response spaces. This multivariate splitting function leads to a natural extension of the relevance scores, which maintains the interpretability.

Thus, interpretability in TSM methods depends in the last instance of relevance scores, which are computed for each input variable (gene expression in our case) by averaging the importance measure (the higher, the better) of each individual tree. Recent studies [61] have concluded that, by means of the averaging of relevance technique, RF could deliver an unreliable importance measure in certain situations, such as classification problems, where the input space has many categorical variables, favoring those variables with a higher number of categories. Although here, predictor and response variables are continuous, multivariate regression is performed instead of classification, the relevance scores have been validated by studying their distribution along the repeated k-fold cross-validation methodology. Figure 5 shows the top 50 gene relevance distributions, ordered by their mean. The genes found as relevant have a significant predictive impact on the circuits as Fig. 6 documents.

By means of the strategy presented here, many of the problems affecting the analysis of genomic Big Data in a ML framework can be overcome to fully exploit the discovery potential of genomic big data.

**Drugs with a potential new indication for FA**

In order to understand what are the general roles played in the cell by the genes selected as most relevant by the ML algorithm (see Table 3) we carried out an enrichment analysis. The functional landscape revealed by the
analysis include Gene Ontology (GO) Biological processes terms mainly related to cell cycle, specifically to the correct regulation of spindle formation, chromatin condensation, centrosome separation and in general, correct mitotic cell phase transition (see Fig. 7 and Additional File 5 for a detailed description of the terms found). These terms specifically involve processes related to DNA replication, DNA repair and stress response, which suggests that the activity of these genes may potentially impact DNA repair cell ability, by controlling the balance between accumulation of mutations and apoptosis in the cell, which indirectly also impacts on tumor predisposition. Interestingly, the rare diseases most associated with relevant genes included Fanconi anemia, as well as other related diseases such as Baller-Gerold syndrome (OMIM:218600), Ataxia telangiectasia (OMIM:208900), Bloom syndrome (OMIM:210900), Filippi syndrome (OMIM:272440), Congenital aplastic anemia (OMIM:609135), Meier-Gorlin syndrome (OMIM:224690), Seckel syndrome (OMIM:606744, OMIM:210600, OMIM:613676, OMIM:613823, OMIM:614728, OMIM:615807, OMIM:616777, OMIM:617253, OMIM:614851), cutaneous melanoma (OMIM:609048). All these diseases share with FA several of its hallmarks like chromosomal instability condition or tumor predisposition [62–64].

Among the most relevant gene drug targets (Table 3) 8 proteins targeted by approved drugs, NEK2, TOP2A, BIRC5, COXI, GRIN1, RARA, SNAP25 and TYMS can be found, revealing the high potential for therapeutic targets and candidates for drug repositioning in FA (Additional File 4) rendered by the ML strategy applied.
Although a detailed discussion on the nature of the most relevant targets is out of the scope of this manuscript, some of the top scored ones deserve to be reviewed for their potential links to FA.

The most relevant protein, NEK2, is a serine/threonine-protein kinase that regulates mitosis. Its expression rises during S phase and reach its maximum level in late G2 phase, just before mitosis. The protein regulates the correct spindle formation and chromatin condensation, playing a major role in cell cycle [65]. Indeed, DNA damage results in G2 arrest due to the drastically decreasing in NEK2 presence [66]. Indeed, this NEK2 inhibition is dependent of ATM, a protein that, along with ATR, are master controllers of cell cycle and DNA repair, the main pathway deregulated in Fanconi Anemia [67]. NEK2 phosphorylates FANCA, a protein conforming the FA core and highly associated with Fanconi Anemia disease [68]. These associations are in line with the expected results, supporting the robustness and suitability of the methodology presented here for the discovery of genes and new therapeutic targets relevant to diseases, FA in this case.

The protein TOP2A is a topoisomerase, a nuclear enzyme that binds to the DNA and alters its topologic state during transcription. It is associated with the initiation of neoplasms, such as breast and peripheral nerve tumors or Bloom syndrome, as well as with several anaemia disorders (Anaemia due to Adenosine triphosphatase deficiency, Congenital dyserythropoietic anaemia and Congenital aplastic anaemia) [69]. Regarding its connection with DNA repair, TOP2A show a consistent high expression in G2, but it is also highly expressed in late S phase, supporting a role in regulating entry into mitosis [70]. Besides, topoisomerase-1 and 2A gene copy numbers are elevated in patients mismatch repair-proficient tumor samples, suggesting that TOP2A is required to deal with high replication stress [71].

Protein BIRC5, also known as survivin, plays an important role in apoptosis, being involved in pathways such as Apoptosis (hsa04210, hsa04215), Hippo signaling pathway (hsa04390) and specific disease pathways such as Pathways in cancer (hsa05200) and Colorectal cancer (hsa05210). Indeed, several studies demonstrate its association with neoplasia and, specifically with colorectal cancer [72]. Some works suggest that the role of survivin in DNA repair by homologous recombination has a direct impact in cancer [73]. The gene BIRC5 is a member of the inhibitor of apoptosis gene family (IAP), thus its downregulation promotes apoptotic cell death. One of the main mechanisms of apoptosis inhibition is due to its protection of the cell towards the action of caspases. Actually, the mechanism by which the Jak/STAT pathway specifically triggers one of the survival circuits of the apoptosis pathway that eventually results in the disease has previously been described by means of a mathematical model [52].

The protein coded by GRIN1, Glutamate Ionotropic Receptor NMDA Type Subunit 1, directly bind thorough NMDA receptors to their ligands (glutamate in this case) allowing calcium to enter the cell, thus, promoting cell activity and proliferation. Interestingly, some studies associate the deregulation of GRIN1 and other NMDA receptors with tumor formation [74].

TYMS (Thymidylate Synthetase) protein plays a critical role in DNA replication and repair [75]. Mutations in its enhancer region, resulting in an overexpression of TYMS, are associated with several cancers and response to chemotherapy [76]. Interestingly, chemotherapeutic agents targeting TYMS, and reducing its expression, have grade 1 anemia as secondary effects, suggesting that deleterious mutations in this gene may produce anemia [77]. Some authors have described that HDAC inhibits both TYMS and BIRC5 (one of the most relevant proteins found by our model), suggesting an indirect relation between both proteins [78]. But not only with BIRC5, a recent study showed a non-canonical interaction between TYMS and FANCD2, a protein belonging to FA pathway [79].

Gene COXI (Mitochondrially Encoded Cytochrome C Oxidase I) codes for the subunit 1 of Cytochrome C oxide, the component of the respiratory chain that catalyzes the reduction of oxygen to water. Defects in this gene are associated with Acquired Idiopathic Sideroblastic Anemia (ORPHA75564), a disease that affects bone, bone marrow and myeloid tissues, phenotypes also present in Fanconi Anemia. COX enzymes have a role in response to oxidative stress, COX-1 is believed to play a constitutive housekeeping role [80] and its inhibition induce apoptosis and lead to Prostaglandin production induced by ionizing radiation [81]. In line with this, it has recently been demonstrated that downregulation of COXI stimulates mitochondrial apoptosis through NH-kB signaling pathway [82].

RARA (Retinoic Acid Receptor alpha) protein is involved in regulation of several cell processes, including cell differentiation, apoptosis and transcription of clock genes. Mutations in RARA gene, mostly resulting in fusion genes, are associated with abnormality of blood forming tissues, leukemias and deregulate genes involved in DNA repair [83]. Recent works have demonstrated in Escherichia coli that rarA, via its gap creation activity, generates substrates for post-replication repair pathways, including homologous recombination and translesion DNA synthesis [84], both DNA repair pathways are involved in FA disease mechanism.

With respect to the 81 drugs targeting the most relevant genes, 55 of them have a description or indication provided by DrugBank, and 28 are already approved as a therapeutic option. Of these, 37 (67.27%) drugs are indicated for cancer treatment (including breast and colorectal cancer, but mostly, leukemias), most of them have antineoplastic effects (23, 38.33%), including chemotherapeutic agents. The
remaining drugs are indicated for a variety of conditions, including infections (viral or bacterial), hyperten-
sion, neuropathies, Alzheimer, schizophrenia or rheuma, acting as anti-inflamatory, antipsychotic, antibacterial or antiviral. Most of the obtained drugs impact in the ability of the cell to perform correct replication and division.

The availability of in vivo and in vitro models for FA [68, 85–87] opens the door to validations of some of these drugs.

**Future directions**

We have demonstrated that the use of circuit activities with a functional meaning in the context of MOR can efficiently discover proteins with an influence over hallmarks of the disease. When these proteins are targets of known drugs, they are potential candidates for repurposing. Actually, systems biology inspired approaches have demonstrated to be superior to conventional reductionistic approaches for drug discovery [88], and especially for drug repurposing [89, 90]. However, training the system with expression in normal tissue is a quite general approach than could be complemented with other potentially interesting data. For example, the Connectivity Map [91] contains 1 million profiles of cell lines treated with different drugs and has been successfully used for drug repurposing using network analysis [92].

On the other hand, there is an extraordinary activity in deep neural networks in the field of bioinformatic applications [93–96], which opens the possibility of developing interpretable deep models in the near future.

**Conclusions**

We have demonstrated how a mechanistic model, which provide a definition of cell functionalities and outcomes that account for the phenotype of the disease, can be used in combination with ML methods and genomic big data available to discover proteins that might have influence over such disease-related cell functionalities and, most likely, on the phenotype of the disease. Depending on the specific molecular mechanism of the disease and the type of influence, the molecules found can be considered therapeutic targets.

Building an interpretable model makes possible understanding how the model learns and, consequently, a disease-centric learning framework can be built. In this way, many of the problems affecting the analysis of genomic data in a ML framework can be overcome to fully exploit the discovery potential of such Big Data.

A gene expression microarray study to identify differences at the transcription level in bone marrow cells between normal volunteers and FA patients [30] was downloaded from GEO (GSE16334) and used to check the performance of the expanded FA disease map model in a real scenario.

Gene expression data from 53 non-diseased tissue sites across nearly 1000 individuals, more than 11,000 samples and about 20,000 gene expression measurements each, were downloaded from the GTEx Portal [98] (GTEx Analysis V7; dbGaP Accession phs000424.v7.p2).

Genes that are target of approved drugs were taken from the DrugBank [35] database (Version 5.1.2). A total of 965 known drug target (KDT) genes targeted by a total of 7122 drugs were considered in this study (see Additional File 1). Some of these genes may potentially affect the whole FA pathway or some of their circuits, affecting in consequence, to the cell functionalities triggered by the affected circuits.

**RNA-seq data processing**

After constructing the gene expression matrix for all samples, the following pipeline was applied: 1) Trimmed mean of M values (TMM) normalization (edgeR package) [99] was applied followed by a 2) Logarithm transformation (apply log(matrix+1)), then 3) Truncation by the quantile 0.99 (all values greater than quantile 0.99 are truncated to this value, all values lower than quantile 0.01 are truncated to this other value) and finally 4) Quantiles normalization (preprocessCore package) [100].

**Mechanistic model of cell functionality**

The normalized gene expression data was rescaled from the range of variation to 0–1 interval range \([\max(matrix)=1, \min(matrix)=0]\). The Hipathia method [21], as implemented in the Hipathia Bioconductor package [48], was used to estimate signaling circuit activities within the expanded FA pathway from the corresponding normalized gene expression values. The Hipathia method uses a Wilcoxon test was used to assess differences in pathway activity between controls and FA samples [21].

**Machine learning**

Here, a Multi-Output Random Forest (MORF) regressor that predicts the circuit activity across the whole disease pathway has been implemented using the scikit-learn general Machine Learning library [101]. In the learning framework used, the multiple dependent variables that conform the disease environment are modeled in a “all at once” fashion, i.e. each signaling circuit activity in the expanded FA pathway is a target/output variable, whereas each expression value of a KDT gene is an input (Multiple Input Multiple Output). In order to find a “quasi-optimal” set of hyperparameters for our MORF
model, we have implemented an optimization strategy on
top of scikit-learn [101] and hyperopt [102]. Since the best
hyperparameters to fit the data are problem-dependent
[103], the hyperparameter space is explored by means of
the TPE [59] method, where each choice of hyperpara-
meters is a “configuration” in the original algorithm. A glo-
bal R² score averaged across a K-fold cross-validation
partition of the data (k = 10) is used as objective function.
Finally, to evaluate the performance of the model in an un-
biased way, the previously found optimal hyperparameters
were fixed and a repeated (N = 10) k-fold cross-validation
is performed.

The same cross-validation can be used to obtain a dis-
tribution of the relevance values that can be used to set
a threshold beyond which the relevance values obtained
by the ML keep their positions in the rank of relevance
(have a stable value).

Enrichment analysis of most relevant genes
Those genes with a relevance confirmed by the cross-
validation procedure were considered relevant and were
used to perform an enrichment analysis to evaluate their
possible impact on the circuits of the FA pathway
triggering FA hallmarks. An enrichment analysis was
performed by using enrichR algorithm using GO Bio-
logical Processes as well as Rare Diseases with AutoRIF
(Automatic Reference into Function) and GeneRIF (Gene
Reference into Function) from ARCHS4 mining of pub-
licly available data tool to predict enrichment in rare dis-
ases terms [104–106].

Additional files

Additional file 1: Table S1. All gene drug targets studied obtained
from DrugBank database version 5.1.2, ranked by their relevance obtained
from MORF modelling. First column: gene name; second column: gene
symbol; third column: Entrez ID; fourth column: relevance; fifth column:
DrugBank ID of the drugs targeting the gene. (XLS 200 kb)

Additional file 2: Table S2. Genes in the KEGG FA pathway (hsa03460).
First column: gene name; second column: KEGG ID; third column: gene
symbol; fourth column: ENSEMBL ID; fifth column: OMIM ID. (DOCX 19 kb)

Additional file 3: Figure S3. Distribution of circuit activities in the FA
KEGG pathway. Distribution of activities in the seven circuits of the FA
KEGG pathway observed in the comparison between healthy and FA
bone marrow cells. (TIF 218 kb)

Additional file 4: Table S4. Drugs targeting most relevant genes
(relevance> 0.005) in Fanconi Anemia extended pathway, obtained from
DrugBank database. First column: DrugBank ID; second column: drug
name; third column: drug description; fourth column: drug status; sixth
column: drug Indication. (XLSX 69 kb)

Table 4 Fanconi Anemia ORPHANET (ORPHA:84) database affected genes

| GENE NAME                                      | SYMBOL | ENTREZ ID | ENSEMBL ID | OMIM  |
|-----------------------------------------------|--------|-----------|------------|-------|
| Fanconi Anemia complementation group F         | FANCF  | 2188      | ENSG00000183161| 603,467|
| Fanconi Anemia complementation group C         | FANCC  | 2176      | ENSG00000158169| 227,645|
| Breast cancer type 2 susceptibility protein    | BRCA2  | 675       | ENSG00000139618| 114,480|
| Breast cancer type 1 susceptibility protein    | BRCA1  | 672       | ENSG0000012048| 113,705|
| Fanconi Anemia complementation group E         | FANCE  | 2178      | ENSG00000112039| 600,901|
| RAD51 recombinase                              | RAD51  | 5888      | ENSG00000051180| 114,480|
| Fanconi Anemia complementation group D2        | FANCD2 | 2177      | ENSG0000014555| 227,646|
| Fanconi Anemia complementation group M         | FANCM  | 57,697    | ENSG00000187790| 609,644|
| DNA repair protein RAD51 homolog 3             | RAD51C | 5889      | ENSG00000108384| 602,774|
| Ubiquitin-conjugating enzyme E2 T              |UBE2T   | 29,089    | ENSG00000077152| 610,538|
| Fanconi Anemia complementation group B         | FANCB  | 2187      | ENSG00000181544| 300,514|
| Fanconi Anemia complementation group G         | FANCG  | 2189      | ENSG000000221829| 602,956|
| Fanconi Anemia complementation group I         | FANCI  | 55,215    | ENSG00000140525| 609,053|
| Fanconi Anemia complementation group L         | FANCL  | 55,120    | ENSG00000115392| 608,111|
| partner and localizer of BRCA2                | PALB2  | 79,728    | ENSG00000083093| 114,480|
| SLX4 structure-specific endonuclease subunit   | SLX4   | 84,464    | ENSG00000188827| 612,278|
| Ring finger and WD repeat domain 3             | RFWD3  | 55,159    | ENSG00000168411| 614,151|
| BRCA1 interacting protein C-terminal helicase 1| BRIP1  | 83,990    | ENSG00000136492| 114,480|
| ERCC excision repair 4, endonuclease catalytic subunit | ERCC4 | 2072      | ENSG00000175595| 133,520|
| Mitotic arrest deficient 2 like 2              | MAD2L2 | 10,459    | ENSG00000116670| 604,094|
| X-ray repair cross complementing 2             | XRCC2  | 7516      | ENSG00000196584| 600,375|
| Fanconi Anemia complementation group A         | FANCA  | 2175      | ENSG00000187741| 227,650|
Additional file 5: Table S5. Enrichment analysis of the most relevant genes. First column: term detected in the enrichment analysis; second column: overlap; third column: p-value; fourth column: adjusted p-value; fifth column: Z score; sixth column combined score; seventh column genes annotated to the term. (XLSX 199 kb)

Abbreviations
FA: Fanconi Anemia; GO: Gene Ontology; KDT: Known Drug Targets; KEGG: Kyoto Encyclopedia of Genes and Genomes; ML: Machine Learning; MDR: Multi-Output Regression; MORF: Multi-Output Random Forest; RF: Random Forest; SOCR: Single Output Regression; TMM: Trimmed mean of M values; TPE: Tree of Parzen Estimators; TSx: Tree-structured methods

Acknowledgements
Not applicable.

Authors’ contributions
ME has performed the data collection and the analysis, MPC has collaborated in the analysis of the data and the discussion, CL has carried out the machine learning computations and JD has conceived the work and wrote the manuscript.

Funding
This work is supported by grants SAF2017–88908-R from the Spanish Ministry of Economy and Competitiveness and “Plataforma de Recursos Biomoleculares y Bioinformáticos” PT17/0009/0006 from the ISCIII, both co-funded with European Regional Development Funds (ERDF) as well as H2020 Programme of the European Union grants Marie Curie Innovative Training Network “Machine Learning Frontiers in Precision Medicine” (MLFPM) (GA 813533) and “ELIXIR-EXCELerate fast-track ELIXIR implementation and drive early user exploitation across the life sciences” (GA 767559).

Availability of data and materials
The datasets analyzed during the current study are available in the GEO repository (accession: GSE16334) [https://www.ncbi.nlm.nih.gov/geo/], GTEx portal (dbGaP accession phs000424.v7.p2) [https://www.gtexportal.org/repository (accession: GSE16334)] [https://www.ncbi.nlm.nih.gov/geo/], GTEx The datasets analyzed during the current study are available in the GEO repository (accession: GSE16334) [https://www.ncbi.nlm.nih.gov/geo/], GTEx

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
1Clinical Bioinformatics Area. Fundación Progreso y Salud (FPS). CDCA, Hospital Virgen del Rocio, 41013 Sevilla, Spain. 2Bioinformatics in Rare Diseases (BiER). Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER). FPS, Hospital Virgen del Rocio, 41013 Sevilla, Spain. 3INB-ELIXIR-es, FPS, Hospital Virgen del Rocío, 41013 Sevilla, Spain.

Received: 6 May 2019 Accepted: 25 June 2019 Published online: 02 July 2019

References
1. Khavjouian A, Quackenbush J, Thompson JF. What would you do if you could sequence everything? Nat Biotechnol. 2008;26(10):1125–33.
2. Mardis ER. DNA sequencing technologies: 2006–2016. Nat Protoc. 2017;12(2):2213.
3. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. Big data: astronomical or genomic? PLoS Biol. 2015;13(7):e1002195.
4. Todoradji E. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44.
5. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, Hanel D, Gardner M, Gupta A, Hotchkiss R. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci. 2018;115(45):11591–6.
6. Beynordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, Van Der Laak JA, Hemmes M, Manson QF, Balkenhol M. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA oncology. 2017;318(22): 2199–210.
7. Ding DS, Liu Y, Burlina P, Xu X, Bressler NM, Wong TY. AI for medical imaging goes deep. Nat Med. 2018;24(5):539.
8. Madani A, Amato R, Mofrad M, Amato R. Fast and accurate view classification of echocardiograms using deep learning. NPJ digital medicine. 2018;1(1).
9. Gaublomme JT, Yosef N, Lee Y, Gertner RS, Yang LV, Wu C, Pandolfi PP, Mak T, Satija R, Shalek AKUC. Single-cell genomics unveils critical regulators of Th17 cell pathogenicity. Cell. 2015;163(6):1400–12.
10. Ding J, Condon A, SPNC S. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat Commun. 2018;9(1):2002.
11. Tan J, Ung M, Cheng C, Greene CS. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. In: Pacific Symposium on Biocomputing Co-Chairs. Kohala Coast: World Scientific; 2014. p. 132–143.
12. Liang M, Li Z, Chen T, JMACbs Z. bioinformatics: integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. IEEE/ACM transactions on computational biology and bioinformatics. 2015;12(4):287–37.
13. Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. Bioinformatics. 2016;32(12):1832–9.
14. Way GP, Sanchez-Vega F, La K, Arneja J, Chata LW, Luna A, Sander C, Cherniack AD, Mina M, Cirilo G. Machine-learning detects pan-cancer ras pathway somatic gains in the cancer genome atlas. Cell reports. 2018;22(12):e172–e173.
15. Edurdu J, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, Tress ML. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. Hum Mol Genet. 2014; 23(22):5866–76.
16. Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, Sharan R, Iبدker T. Using deep learning to model the hierarchical structure and function of a cell. Nat Methods. 2018;15(4):290.
17. Carvunis A-R, Iبدker T. Sir of the cell: what biology could learn from the iPhone. Cell. 2014;157(3):534–8.
18. Yu MK, Kramer M, Dutkowski J, Sivas R, Licon K, Kreisberg JF, Ng CT, Krogan N, Sharan R, Iبدker T. Translation of genotype to phenotype by a hierarchy of cell subsystems. Cell systems. 2016;2(2):77–88.
19. Amadoz A, Sebastan-Leon P, Vidal E, Salavert F, Dopazo J. Using activation status of signaling pathways as mechanism-based biomarkers to predict drug sensitivity. Sci Rep. 2015;5:18494.
20. Cubuk C, Hidalgo MR, Amadoz A, Rian K, Salavert F, Pujana MA, Mateo F, Herranz C, Carbonell-Caballerio J, Dopazo J, et al. Differential metabolic activity and discovery of therapeutic targets using summarized metabolic pathway models. NPJ Systems Biology. 2015;1(17).
21. Hidalgo MR, Cubuk C, Amadoz A, Salavert F, Carbonell-Caballero J, Dopazo J. High throughput estimation of functional cell activity reveals disease mechanisms and predicts relevant clinical outcomes. Oncotarget. 2017;8(35):160–70.
22. Cubuk C, Hidalgo MR, Amadoz A, Pujana MA, Mateo F, Herranz C, Carbonell-Caballero J, Dopazo J. Gene expression integration into pathway modules reveals a pan-cancer metabolic landscape. Cancer Res. 2018;78(21):6599–72.
23. Fey D, Halasz M, Drediao D, Kennedy SP, Hastings JF, Rauch N, Munoz AG, Pilkington R, Fischer M, Westermann F, et al. Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. Sci Signal. 2015;8(408):ra130.
24. Cubuk C, Hidalgo MR, Amadoz A, Kian K, Salavert F, Pujana MA, Mateo F, Herranz C, Carbonell-Caballero J, Dopazo J. Differential metabolic activity and discovery of therapeutic targets using summarized metabolic pathway models. NPJ Systems Biology. 2015;1(17).
25. Razzoli M, Frontrini A, Guney A, Mondini E, Cubuk C, Katz LS, Cero C, Bolan PJ, Dopazo J, Vidal-Puig A. Stress-induced activation of brown adipose tissue prevents obesity in conditions of low adaptive thermogenesis. Proc Natl Acad Sci. 2015;112(19):E2753–63.
26. Ferreira PG, Ciriello G, Ideker T. Siri of the cell: what biology could learn from the iPhone. Cell. 2014;157(3):534–8.
27. Taniguchi T, D’Andrea AD. Molecular pathogenesis of Fanconi anemia: recent progress. Blood. 2006;107(11):4223–7.
28. Nakanishi K, Yang Y-G, Pierce AJ, Taniguchi T, Digweed M, D’Andrea AD, Wang Z-Q, Jin A. Human Fanconi anemia monoubiquitination pathway promotes homologous DNA repair. Proc Natl Acad Sci. 2005;102(4):1110–5.
29. Walden H, Deans AJ. The Fanconi anemia DNA repair pathway: structural and functional insights into a complex disorder. Annu Rev Biophys. 2014;43:257–78.

30. Vanderweer SM, Swahn J, Olson S, Rethbun RK, Harrington C, Yates J, Kleeve W, Anderson DC, Anur P, Pereira EF, et al. TF8-dependent TGF-alpha overexpression in Fanconi anemia group C cells. Blood. 2009;114(16):2590–8.

31. Minguillón J, Surralles J. Therapeutic research in the crystal chromosome disease Fanconi anemia. Mutat Res. 2018;836:104–8.

32. Sissen S, Cissamain D, Doms M, Piccett A. Orphan drugs for rare diseases. Drugs. 2012;72(11):1437–43.

33. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004;3(6):573.

34. Delavan B, Roberts R, Huang R, Bao W, Tong W, Liu Z. Computational drug repositioning for rare diseases in the era of precision medicine. Drug Discov Today. 2017.

35. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajedi T, Johnson D, Li C, Sayerda Z. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research. 2017;46(D1):D1074–82.

36. Rani J, Shah AR, Ramachandran S. Pubmed miner: an R package with text-mining algorithms to analyze PubMed abstracts. J Bioc. 2015;8(4):671–82.

37. Tomida J, Takata K, Lange SS, Schibler AC, Yousefzadeh MJ, Bhetawal S, Dent SY, Wood RD. REV7 is essential for DNA damage tolerance via two REV7 binding sites in mammalian DNA polymerase ζ. Nucleic Acids Res. 2018;45(2):1000–11.

38. Elia AE, Wang DC, Willis NA, Boardman AP, Hajdu I, Adeyemi RO, Lowrey E, Gygi SP, Scorza R, Eldridge SJ. RFWD3-dependent ubiquitination of RPA regulates repair at stalled replication forks. Mol Cell. 2015;60(2):280–93.

39. Tambini CE, Spink KG, Ross CJ, Hill MA, Thacker J. The importance of XRCC2 in mammalian DNA-damage tolerance in response to UV irradiation. FEBS J. 2019.

40. Walden H, Deans AJ. The Fanconi anemia DNA repair pathway: computational biology and bioinformatics. Wiley Interdisciplinary Reviews Data Mining. 2012;2(6):493–507.

41. Banfield RE, Hall LO, Bowyer KW, Kegelmeyer WP. Intelligence m: a comparison of decision tree ensemble creation techniques. IEEE transactions on pattern analysis. 2007;29(1):173–80.

42. Qi Y. Random forest for bioinformatics: Ensemble machine learning. Boston: Springer; 2012. p. 307–23.

43. Batselustar R, De Andres SA. Gene selection and classification of microarray data using random forest. BMC bioinformatics. 2006;7(1):3.

44. Wang Y, Goh W, Wong L, Montana G. Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes. BMC bioinformatics. 2013;14(1):56.

45. Bergstra JS, Bardenet R, Bengio Y. Kégl B. Algorithms for hyper-parameter optimization. In: Advances in neural information processing systems. 2011: 2546–54.

46. Segal MR. Tree-structured methods for longitudinal data. J Am Stat Assoc. 1992;87(418):407–18.

47. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC bioinformatics. 2007;8(1):25.

48. Balteke C, Patirolu T, Gumruk F. Fanconi Anemia and Ataxia telangiectasia in siblings who inherited unique combinations of novel FANCA and Ataxia telangiectasia mutations. J Pediatr Hemat Oncol. 2019;41(3):243–6.

49. Moniz L, Dutt P, Haider N, Stambolic V. Nek family of kinases in cell cycle, checkpoint control and cancer. Cell Div. 2011;6(1):18.

50. Fletcher L, Cerniglia G, Ngig EA, Yen TJ, Muschel RJ. Inhibition of centrosome separation after DNA damage: a role for Nek2. Radiat Res. 2004;162(2):128–35.

51. Mi J, Guo C, Brautigal DL, Lamer JM. Protein phosphatase-1 regulates centrosome splitting through Nek2. Cancer Res. 2007;67(3):1082–9.

52. Dong H, Nebert DW, Bruford EA, Thompson DC, Joenje H, Vasiliiu V. Update of the human and mouse Fanconi anemia genes. Hum Genomics. 2015;9(1):32.

53. Leo AD, Desmedt C, Bartlett JMS, Piette F, Ejlertsen B, Larsimont D, Ro C, Thiboutot S, et al. Integrative histologic and bioinformatics analysis of BIRC5/Survivin expression in Oral squamous cell carcinoma. Int J Mol Sci. 2017;18(9):1886.

54. Fuku T, Tsukiyomt K, Srinath B, Bhatnagar B, Burke RA, Berman SH, Bleucu P, Riaz N, HUO Y. Compromised BRCA1–PALB2 interaction is associated with breast cancer risk. Oncogene. 2017;36(20):4161.

55. Conde M, Michen S, Wiedermuth R, Klinkn B, Schröck E, Schackert G, Tenne A. Chromosomal instability induced by increased BIRC5/Survivin levels affects tumorigenicity of glioma cells. BMC Cancer. 2017;17(1):889.

56. Gorska-Ponikowska M, Perricone U, Kuban-Jankowska A, Lo Bosco G, Barone G. 2-methoxyestradiol impacts on amino acids-mediated metabolic repogramming in osteosarcoma cells by its interaction with NMDA receptor. J Cell Physiol. 2017;232(11):3030–49.

57. Kotoula V, Krikelis D, Karavasilis V, Koletsa T, Eleftheraki AG, Televantou D, Anagnostou K, Theodoridou M, et al. 2-methoxyestradiol reduces the number of 2-neurotrophin-3 expressing neurons in the hippocampus and hyperactivates the NMDA receptors in the dentate gyrus of the rat. Brain Res. 2017;1659:147–56.
synthase (TYMS) enhancer region genotype-directed phase II trial of oral capecitabine for 2nd line treatment of advanced pancreatic cancer. Invest New Drugs. 2011;29(5):1057–62.
78. Bhatia T, Wang J, Morrison DJ, Raetz EA, Burke MJ, Brown P, Carroll WL. Epigenetic reprogramming reverses the relapse-specific gene expression signature and restores chemosensitivity in childhood B-lymphoblastic leukemia. Blood. 2012;119(2):5201.
79. Zhang T, Du W, Wilson AF, Namekawa SH, Andreassen PR, Meeitei AR, Pang Q. Fancd2 in vivo interaction network reveals a non-canonical role in mitochondrial function. Sci Rep. 2017;7:45626.
80. Burdon C, Mann C, Cindrova-Davies T, Ferguson-Smith AC, Burton GJ. Oxidative stress and the induction of cyclooxygenase enzymes and apoptosis in the murine placenta. Placenta 2007; 28(7):724–733.
81. Benitez-Rangel E, García L, Namorado MC, Reyes JL, Guerrero-Hernández A. Ion channel inhibitors block caspase activation by mechanisms other than restoring intracellular potassium concentration. Cell Death & Disease. 2011;2:e113.
82. Ding L, Gu H, Lan Z, Lei Q, Wang W, Ruan J, Yu M, Lin J, Cui Q. Downregulation of cyclooxygenase-1 stimulates mitochondrial apoptosis through the NF-κB signaling pathway in colorectal cancer cells. Oncol Rep. 2019;41(1):559–69.
83. Alcayal M, Meani N, Gelmetti V, Fantozzi A, Fagioli M, Orleth A, Riganelli D, Sebastiani C, Cappelli E, Cescati C, et al. Acute myeloid leukemia fusion proteins deregulate genes involved in stem cell maintenance and DNA repair. J Clin Invest. 2003;112(11):1751–61.
84. Staneje TH, Page AN, Cox MM. DNA flap creation by the RarA/MgsA protein of Escherichia coli. Nucleic Acids Res. 2017;45(5):e2724–35.
85. Parmar K, D’Andrea A, Niedernhofer LJ, Mouse models of Fanconi anemia. Mutat Res. 2009;668(1–2):133–40.
86. Liu G-H, Shi J, Le Q, Gu H, Lan Z, Lei Q, Wang W, Ruan J, Yu M, Lin J, Cui Q. Modelling Fanconi anemia pathogenesis and therapeutics using integration-free patient-derived iPSCs. Nat Commun. 2014;5:4330.
87. Roe P, Bahos R, Lombardo A, Quintana-Bustamante O, Alvarez L, Garate Z, Genovese P, Almarza E, Valer A, Díez B, et al. Targeted gene therapy and cell reprogramming in Fanconi anemia. EMBO Molecular Medicine. 2014;6(9):835–48.
88. Ryall KA, Tan AC. Systems biology approaches for advancing the discovery of effective drug combinations. Journal of cheminformatics. 2015;7(1):7.
89. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. Brief Bioinform. 2015;17(1):1–2.
90. Hurle M, Yang L, Xie Q, Rajpal D, Sanseau P, Agarwal P. Therapeutics: computational drug repositioning: from data to therapeutics. Clinical Pharmacology. 2013;39(4):335–41.
91. Subramanian A, Narayan R, Corcelli SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171(16):1437–52 e1417.
92. Regan-Fendt KE, Xu J, Díez B, Almazán F, Frémeaux-Bacchi V, Santini D, Park W, Caron WE, Payne PRQ, Li F. Synergy from gene expression and network mining (SynGeNet) method predicts synergistic drug combinations for diverse melanoma genomic subtypes. npj Systems Biology and Applications. 2019;5(1):6.
93. Li Y, Huang C, Ding L, Li Z, Pan Y, Gao X. Deep learning in bioinformatics: introduction, application, and perspective in big data era. arXiv:2019;1603.04967.
94. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2017;18(5):851–69.
95. Eraslan G, Avsec Z, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;1.
96. Zou J, Huss M, Abd I, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2018;1.
97. Pavan S, Rommel K, Marquina MEM, Höhn S, Lanneau V, Rath A. Clinical practice guidelines for rare diseases: the orphanet database. PLoS One. 2017;12(1):e0170365.
98. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasc R, Walters G, Garcia F, Young N. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45(6):580.
99. Robinson MD, McCarthy DJ, Smyth GK, EdgeR. a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
100. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185–93.
101. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12(Oct):2825–30.

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

Ready to submit your research? Choose BMC and benefit from:
- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is already in progress.
Learn more biomedcentral.com/submissions