bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Germline genetic variants associated with leukocyte-genes predict tumor recurrence in breast cancer patients

Jean-Sébastien Milanese[1], Chabane Tibiche[1], Jinfeng Zou[1], Pengyong Han[2], Zhi Gang Meng[2,3], Andre Nantel[1], Simon Drouin[1], Richard Marcotte[1,4], and Edwin Wang[2,5*]

1. National Research Council Canada, 6100 Royalmount Ave, Montreal, Canada, H4P 2R2
2. Department of Biochemistry & Molecular Biology, Medical Genetics, and Oncology, University of Calgary, 3330 Hospital Dr. NW, Calgary, Canada, T2N 4N1
3. Chinese Academy of Agricultural Science, No. 12 Zhongguangcun South St., Haidian District, Beijing, 100086, China
4. Rosalind and Morris Goodman Cancer Research Centre, McGill University, Quebec, Canada, McGill University, 1160 Pine Ave W, Montreal, Canada, H3A 1A3
5. Alberta Children's Hospital Research Institute and Arnie Charbonneau Cancer Research Institute, University of Calgary, 3330 Hospital Dr. NW, Calgary, Canada, T2N 4N1

*Corresponding to: E.W. (edwin.wang@ucalgary.ca)
Department of Biochemistry & Molecular Biology, Medical Genetics, and Oncology, University of Calgary, 3330 Hospital Dr. NW, Calgary, Canada, T2N 4N1
Phone: 403-220-6609
Fax: 403-9891-9879

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Germline genetic variants such as BRCA1/2 play an important role in tumorigenesis and clinical outcomes of cancer patients. However, only a small fraction (i.e., 5-10%) of inherited variants has been associated with clinical outcomes (e.g., BRCA1/2, APC, TP53, PTEN and so on). The challenge remains in using these inherited germline variants to predict clinical outcomes of cancer patient population. In an attempt to solve this issue, we applied our recently developed algorithm, eTumorMetastasis, which constructs predictive models, on exome sequencing data to ER+ breast (n=755) cancer patients. Gene signatures derived from the genes containing functionally germline genetic variants significantly distinguished recurred and non-recurred patients in two ER+ breast cancer independent cohorts (n=200 and 295, $P=1.4x10^{-3}$). Furthermore, we found that recurred patients possessed a higher rate of germline genetic variants. In addition, the inherited germline variants from these gene signatures were predominately enriched in T cell function, antigen presentation and cytokine interactions, likely impairing the adaptive and innate immune response thus favoring a pro-tumorigenic environment. Hence, germline genomic information could be used for developing non-invasive genomic tests for predicting patients' outcomes (or drug response) in breast cancer, other cancer types and even other complex diseases.

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Cancer is a process of asexual evolution driven by genomic alterations. A single normal cell randomly acquires a series of mutations that allows it to proliferate and to be transformed into a cancer cell (i.e., founding clone) that then initiates tumor progression and recurrence. In general, cancer recurrence and metastasis are the result of the interactions of multiple mutated genes. New somatic mutations arise and are selected if they confer a selective fitness advantage (e.g., proliferation, survival, etc.) to a founding clone in the context of a pre-existing genomic landscape (i.e., germline genetic variants). Hence, pre-existing germline genetic variants provide a profound constraint on the evolution of tumor founding clones and subclones, and therefore, have a contingent effect on the genetic makeup of tumor and presumably patient outcomes. Family history remains one of the major risk factors that contribute to cancer and recent studies have identified several genes whose germline mutations are associated with cancer. For example, patients suffering from Li-Fraumeni syndrome have an almost 100% chance of developing a wide range of malignancies before the age of 70. Most patients carry a missing or damaged *p53* gene, a tumor suppressor whose activity is impaired in almost 50% of all cancers. Other cancer-predisposition genes include *BRCA1* and *BRCA2*[1,2], which are associated with breast and ovarian cancer, *PTEN*[3], whose mutation results in Cowden syndrome, *APC*, which is linked to familial adenomatous polyposis[4] and the Retinoblastoma gene *RB1*[5]. Two distinct types of multiple endocrine neoplasias are associated with the *RET* and *MEN1*[6] genes while *VHL* alterations result in kidney and other types of cancer[7]. Finally, Lynch syndrome, a form of colorectal cancer, is linked to *MSH2*, *MLH1*, *MSH6*, *PMS2*, and *EPCAM*[8]. Genetic tests based on these highly-penetrant gene mutations have shown their usefulness, but they can explain only a small fraction (5-10%) of patients. Most neoplasms arise and are modulated by the interactions of multiple genes and there is a great diversity of genetic alterations even within tumors of the same subtypes.

Thus far, it is unclear to what extent germline genetic variants affect tumorigenesis, tumor evolution and even clinical outcome. We have previously shown that tumor founding clone mutations are able to predict tumor recurrence[9]. Here, we reasoned that the collective impact of germline genetic variants in cancer patients might largely determine tumorigenesis, evolution and even clinical outcomes. That is, germline genetic variants act in combination with newly acquired somatic mutations to modulate tumorigenesis and tumor recurrence. The combination of germline variants and somatic mutations of each patient predispose to the

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

specific activation of biological/signaling pathways (even phenotypes) that directly impact clinical outcomes of cancer patients. Therefore, the germline genomic landscape of cancer patients might predict disease progression. Yet, thus far, clinical outcome predictions using cancer germline genomic information have been limited to only a few cancer types, or to a limited number of gene[1,2,3,4,5,6,7,8]. The increasing availability of genome sequencing data provide opportunities to develop predictive models that can translate these complex genomic alterations into clinical use.

In this study, we showed that the collective germline genetic variants of breast cancer patients predict tumor recurrence by applying a recently developed method, eTumorMetastasis[9], to 755 breast cancer patients. Further statistical analyses showed that the leukocyte gene expression levels and tumor-infiltrating leukocytes (TIL) fractions within tumors between the two predicted groups were significantly different**.** Germline mutated variants associated with tumor recurrence likely impair the adaptive immune response functions of affected individuals, increasing the susceptibility to relapse. These results highlight the important role of germline genetic variants in tumor evolution and recurrence.

Results

Germline genetic variants predict breast cancer recurrence

To examine if germline genetic variants are able to predict tumor recurrence, we used whole-exome sequencing data (i.e., from the NCI Genomic Data Commons, GDC) of healthy tissues from 755 ER+ breast patients by applying our recently developed method, eTumorMetastasis[9]. ER+ subtype represents ~70% of breast cancer patients, thus, in this study, we used only patient data from this subtype. The demographic table of the breast cancer cohort is represented in Table 1.

We hypothesized that somatic mutations are evolutionary selected to work with the pre-existing germline genetic variants to initiate tumorigenesis and recurrence. This is the underlying concept of eTumorMetastasis. In turn, the model infers that pre-existing germline genetic variants of cancer patients have predictive power for recurrence and clinical outcomes. eTumorMetastasis contains 3 main components: (1) a network-based approach[11,12] to transform functionally genetic variants information on a cancer type specific signaling network; (2) identifying biomarkers via our previously developed method, MSS (Multiple

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Survival Screening)[13] and (3) a better predictive power using our previously developed method by combining biomarkers[14]. The detailed procedure of eTumorMetastasis and network construction were described previously[9]. To apply eTumorMetastasis, briefly we first annotated the functional variants using the germline whole-exome sequencing data of each breast cancer patient (see Methods and Supplementary Methods), constructed an ER+ breast cancer-specific molecular network for recurrence (see Methods), and then mapped the functional germline-variant genes on the recurrence network. Finally, gene signatures (i.e., biomarkers) were obtained using MSS and eTumorMetastasis.

We used the germline genomic information of 200 ER+ breast cancer samples (i.e., training samples) to identify gene signatures (i.e., because eTumorMetastasis identifies network-based gene signatures, we called the gene signatures Network Operational Signatures or NOG signatures), which could distinguish recurred and non-recurred breast tumors. By applying eTumorMetastasis to the germline genomes of 200 patients, we identified 18 NOG signatures (Supplementary Tables 1 and 2) for ER+ breast cancer. Each NOG contains 30 genes and represents a cancer hallmark such as apoptosis, cell proliferation, metastasis, and so on. We have previously shown that multiple gene signatures representing distinct cancer hallmarks could be identified from one training cohort[13]. Furthermore, ensemble-based prediction using multiple gene signatures representing distinct cancer hallmarks significantly improved prediction performance[14]. Thus, we used the 18 NOG gene signatures to construct a NOG_CSS (i.e., NOG-based Combinatory Signature Set) using a testing set of 60 samples based on the method we developed previously[14]. We then used the NOG_CSS to predict the prognosis of ER+ breast cancer patients. As shown in Figure 1 and Table 2, we demonstrated that the germline-derived NOG_CSS significantly distinguished recurred and non-recurred breast tumors in two validations sets: 200 (ER+ Nature-Set, $P=1.4x10^{-2}$), 295 (ER+ TCGA-CPTAC independent set, $P=1.4x10^{-3}$). These results suggest that germline genetic variants are significantly correlated with tumor recurrence and support our hypothesis that the original germline genomic landscape of a cancer patient has a significant impact on clinical outcome.

As a proof of concept and to further demonstrate the constraint given by germline variants onto the tumor development, we used the NOG_CSS and the gene expression of normal tissue of 72 breast cancer patients to predict patients relapse risk (See Methods for details). Samples were assigned in the training or validation set previously defined. The results of this prediction can be found in Table 3. Accuracy for low-risk samples was similar to germline

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variants predictions (88.9% compared to 94.9%) suggesting that the impact from germline variants is also reflected in gene expression and correlates with our hypothesis that gene expression and tumor development are affected directly from germline predispositions. Strikingly, the accuracy obtained for high-risk samples with gene expression data was much better than what we obtained using germline variants (66.7% compared to 21.0%), suggesting that gene expression is a better predictor of recurrence for high-risk patients or that high-risk patients might possess a more complex somatic landscape not captured solely by germline mutations.

To compare the prediction performance of the NOG_CSS with clinical factors, we conducted relapse-free survival analysis of clinical factors using Cox proportional hazards regression model. The best p-value (i.e., $P=1.0x10^{-2}$., log-rank test) using covariate models (Supplementary Table 3) was not better than the one derived from the germline NOG_CSS ($P=1.4x10^{-3}$). These results suggest that gene signatures derived from germline genomic information have a better predictive performance than clinical factors.

Finally, we also assessed the number of functional germline genetic variants in all genes or genes specifically expressed in leukocytes as well as the number of genes harboring germline genetic variants for both predicted risk group. Student T-tests revealed a significant difference for all the comparisons ($1.29x10^{-13}$, $8.24x10^{-16}$ and $1.14x10^{-5}$, respectively) with functional germline genetic variants in leukocyte expressed genes being the most indicative distinction. All distributions are highlighted in Figure 2. A higher germline functional mutation count for high-risk group suggests once again that germline variants have a profound impact on tumor development and therefore, recurrence.

Predictive germline genetic variants potentially impair both the innate and adaptive immune system of breast cancer patients

To further understand why germline genomic landscapes of cancer patients are predictive for tumor recurrence, we ran enrichment analyses for genes present in the NOG signatures of breast cancers using DAVID[15]. Interestingly, most genes were enriched in immune- or cell proliferation-related biological pathways and Gene Ontology (GO) terms (Supplementary Tables 5).

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Thus, we hypothesized that recurred patients have more functionally inherited variants in immune system related genes than non-recurred patients. To test this hypothesis, we compared gene expression for leukocytes metagenes between predicted recurred and non-recurred patients from tumor transcriptomes. The leukocytes metagene list was obtained from a recent study[16]. Student's T-tests between both groups revealed a significant difference for myeloid-derived suppressor cells (MDSCs), effector memory CD8 T cells (E-Memory CD8+ T cells), activated dendritic cells (DCs+), activated CD8 T cells (CD8 T cells+), T follicular helper cells (Tfh), monocytes (Monos), memory B cells and activated B cells (B cells+)($P=1.99x10^{-3}$, $P=4.03x10^{-3}$, $P=6.67x10^{-3}$, $P=2.10x10^{-2}$, $P=2.30x10^{-2}$, $P=3.78x10^{-2}$, $P=4.37x10^{-2}$ and $P=4.46x10^{-2}$, respectively). To a similar extent, we also analyzed TILs fractions to see if these were different between predicted groups (CIBERSORT LM22, see Methods)[16,17]. Student's T-tests revealed a significant difference in TILs fractions for gamma delta T cells (γδ T cells) , resting natural killer cells (NK cells-), resting mast cells (MCs-) and CD8+ T cells ($P=3.14x10^{-2}$, $P=4.29x10^{-2}$, $P=4.97x10^{-2}$, $P=8.21x10^{-3}$, respectively). A better representation of leukocytes gene expression and TILs fractions between the predicted groups are shown in Figure 3 and the complete abbreviation lists can be found in Supplementary Tables 6-7. Overall, these results suggest that germline genetic variants of cancer patients could directly influence gene expression and alter immune system functions, cell division and the immune tumor microenvironment (TME). Modulation of these pathways would then affect recurrence and patient outcome.

To further investigate the predictive power of variants in leukocytes-expressed genes, we re-ran eTumorMetastasis[9] pipeline using only functional germline variants in leukocytes-expressed genes (Supplementary Methods). Interestingly, we weren't able to obtain enough significant gene sets to extract a gene signature proposing that leukocytes variants only provides partial information and that the complete germline mutational landscape is more representative.

Discussion

With the recent advances in new generation sequencing, genetic testing and patients monitoring has increased drastically. Yet, it remains a challenge to stratify patients in order to provide an optimal treatment route. Here, we developed a risk classification method using germline genomic variants to predict clinical outcomes and demonstrate that these germline variants shape tumor evolution and recurrence. The enrichment analysis of the NOG

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

signatures derived from germline genetic variants suggest that recurred patients differently regulate signaling pathways associated with immune responses (such as inflammation and cell adhesion).

Breast cancer patients with no lymph node involvement often undergo unnecessary adjuvant chemotherapy treatment (70-80% of patients). In fact, toxic therapies are given to most women with early-stage breast cancer from which 60–75% will not receive any benefit, but instead will experience only side effects[19]. Therefore, biomarkers identification to accurately stratify low-risk breast cancer patients who will not benefit from adjuvant chemotherapy is essential. We showed that germline variants enabled to identify low-risk patients with an accuracy as high as 94.9%, suggesting that inherited variants of breast cancer patients are useful in clinical applications.

Comparison of germline variants and affected genes between the two predicted groups indicates that these variants are predisposing to cancer. Potentially, a significantly higher number of functional variants could lead to a greater number of impaired proteins which would create an imbalance in signaling pathways, favoring tumor development and recurrence. As mentioned above, a functionally mutated gene does not always translate into an expressed loss-of-function protein but the difference between both groups suggest that nonetheless, the greater number of germline variants increases patient susceptibility to relapse. Moreover, we found that leukocyte genes harbored a greater number of germline genetic variants in the predicted high-risk group. These germline genetic variants likely impede the immune system, leading to a more favorable environment for tumor developments.

Our study suggests that tumor recurrence is predicted by the germline genomic landscapes of cancer patients. We found that germline variants in genes regulating cell division, immune cell infiltration and T cell activities are predominately predictive for tumor recurrence. More specifically, mutations in the antigen processing and presentation pathway could impair neoantigens presentation at the surface of cancer cells so that T cells are no longer able to recognize tumor cells, allowing them to evade immune detection. Furthermore, mutations in cell division process could introduce a higher number of somatic mutations during cell division directly promoting tumor development. Activation of Wnt pathway can also block the infiltration of immune cells within tumors[20]. TILs fractions analyses also reveal strong

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

correlation with germline prediction and differential expression in MDSCs, CD8+ T cells, DCs, Tfh cells, monocytes and B cells (Fig. 3A). Aside from memory B cells, all other TILs were enriched in the predicted low-risk group. B cells have been shown to secrete pro-tumorigenic factors (e.g., angiogenesis, tumor growth) and also to inhibit the anti-tumor immune response via cytokines[21-23]. DCs are well known for their role in antigen presentations and in initiating an adaptive immune response[24]. Tfh cells have been shown to favor an adaptive immune response via the B cell chemoattractant CXLC13 in breast cancer[25]. Along with E-memory CD4 T cells, E-memory CD8 T cells possess a key role in the immune response and tumor infiltration. Patient survival has been directly correlated with CD8 T cells infiltration. Multiple mechanisms are used by cancer cells to escape immune responses such as altering cytokines and chemokines attraction to create a non-inflammatory environment which, in turn, inhibits T cells infiltration[26,27]. Monocytes and MDSCs have largely been associated with tumor recurrence in the literature. Monocytes differentiation into tumor-associated macrophages (TAMs) promotes anti-immunity signals such as angiogenesis and growth factors resulting in a TME favoring cancer cell proliferation. However, there have been some reports indicating that a nonclassical monocyte subtype, patrolling monocytes, reduces tumor recurrence by recruiting NK cells[28,29]. Monocytes can also differentiate into pro-inflammatory M1 macrophages aiding the adaptive immune response. A recent study has also shown that TNFα secreted by T cells induces emergency myelopoiesis resulting in an increase in MDSCs in mice[30]. TNFα secretion by T cells could be a regulation mechanism induced by the adaptive immune response once a certain concentration of T cells has infiltrated the tumor. This point could explain the higher expression numbers for MDSCs in predicted low-risk samples.

A significant difference was also seen in TILs cell fractions of γδ T cells, CD8 T cells, NK cells- and MCs (Fig. 3B) between both predicted groups. As mentioned above, CD8 T cell tumor infiltration is crucial for an optimal immune response; these cells were present in greater numbers in the predicted low-risk group. Gamma delta T cells (γδ T cells) are known to have dual effects, capable of exerting both pro- or anti-tumor response depending on their subtype[31]. γδT1, γδT-APC and γδTfh subtypes all possess antitumor activity such as secreting chemoattracting chemokines (i.e. CXLC13), antigen presentation and antibody-dependent cell-mediated cytotoxicity towards cancer cells[32]. Tumor-infiltrating mast cells (MCs) can either emit pro- or anti-tumor signals. Specifically, in breast cancer, MCs are linked with pro-angiogenic factors such as inflammation[33,34] reflecting a higher MCs count in the predicted

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

high-risk group. Finally, NK cells have cytotoxic abilities and a greater number in tumors is indicative of a good prognosis[35,36]. Overall, germline genetic variants identified in the high-risk group could directly modulate gene expression in immune cells, resulting in a weakened adaptive immune response.

Our understanding of the biology mediating recurrence is limited. Germline variants of cancer patients could affect the activity of the immune system in TMEs. For example, germline-encoded receptor variants were shown to trigger innate immune response in cancer patients[37]**.** In addition, lung cancer patients with a germline mutation in Nrf2 have a good prognosis because these variants regulate the inflammatory status and redox balance of the hematopoietic and immune systems of cancer patients[38]. In prostate cancer, patients with a germline variant of the ASPN D locus are associated with poorer outcomes[39]. These studies, including our own, highlight the impacts of germline variants on tumor recurrence and provide a rationale to further study the effect of germline genomic landscapes on clinical outcomes of carcinogenesis.

Good accuracy obtained using normal tissue RNA prediction shows that germline variants directly influence gene expression and consequently, tumor development. A higher accuracy for the high-risk group also highlights that gene expression holds a better predictive power than genome sequencing. These results are not surprising considering that gene expression integrates more information than gene-coding mutations alone (e.g, gene regulation). Even the most damaging functional mutation in a gene not expressed would have no impact on the phenotype. However, gene expression usually requires a biopsy. Therefore, exome-sequencing of normal tissue provides a much more convenient and less invasive method for clinical purposes. We also note that this analysis suffers from a small sample size and should be further explored in the future.

In all, these results suggest that germline variants potentially alter the immune system and the immune TME which in turn stimulate tumor recurrence and ultimately, affect patient outcome. Traditionally, germline variants have been largely ignored in the cancer genomic community; for example, most of the cancer genomic studies including the GDC and The Cancer Genome Atlas (TCGA) have often focused only on somatic mutations while germline mutations were filtered out before formal analysis of tumor genome sequencing data. The demonstration that germline exome sequencing data can predict cancer patients' outcomes

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

suggests that non-invasive genomic tests of cancer patients could be devised to determine cancer prognosis and inform treatment decisions. Genome-wide germline genetic variants can be easily identified by genome/whole-exome sequencing of liquid biopsies such as blood or saliva samples. Prognostic prediction using a patient's germline genomic landscape opens up the possibility of assessing cancer patients' risk of recurrence in a non-invasive manner, which allows for a better forecasting of cancer recurrence in a quick, convenient and minimally invasive manner.

Methods

Exome data processing

We obtained whole-exome sequencing data of breast cancers from the GDC: ER+ breast cancer, a training set of 200 samples, a testing set of 60 samples and two independent validation sets of 200 and 295 samples (TCGA-Nature and TCGA-CPTAC, respectively, Supplementary Table 4). Raw sequence reads from healthy samples of cancer patients were processed using GATK[40] pipeline and the method described previously[9]. Variant calling was then performed using Varscan2[41].

Transcriptome data processing

Normal tissue RNA-seq is less accessible on the GDC than tumor RNA-seq data. Out of 755 samples in our dataset, we were only able to find 72 samples from which normal tissue RNA-seq was available. FPKM values for each sample were downloaded and then normalized using z-score normalization. Each sample was then assigned to our previously defined training and validation set (23 and 49, respectively).

Germline variant identification

To determine germline variants, we used variant allele frequencies (VAFs) between the tumor and healthy samples. We defined homozygous germline variants if the VAF in the healthy samples was >=90. For heterozygous germline variants, we used the VAF cutoffs between 45 and 65% in healthy samples. Only germline functional variants were retained for downstream analysis.

Germline NOG signature identification

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To identify NOG signatures using the functional mutated genes of breast cancer patients' germline genomes, we followed the eTumorMetastasis[9] method. Briefly, we constructed a breast-specific recurrence network based on the methods described previously[9]. For each patient, we used its germline functionally mutated genes as seeds on the breast cancer-specific recurrence network to perform network propagation and then identify NOG signatures.

Transcriptomic normal tissue prediction

Like mentioned above, each sample was assigned to our previously defined training and validation set (23 and 49, respectively). Accuracy and recall rate were obtained using a similar approach than with the eTumorMetastasis[9] method. For all 18 NOG signatures previously identified with genome sequencing, we calculated centroids values for each gene between both groups (high- and low-risk) in the training set. In this case, centroids values were obtained from gene expression values instead of network propagation scores. For each sample in our validation set, each sample was classified to its specific group based on Pearson correlation with centroids from both groups. We built a NOG_CSS using the same cutoffs obtained from genome sequencing. Prediction accuracy and recall rate for validation samples can be found in Table 3.

Leukocytes metagene expression and cell fractions

Leukocytes metagene expression derived from tumor RNA-seq data were obtained from The Cancer Immunome Atlas (TCIA)[16] and were applied z-score normalization. In total, scores for 29 leukocytes metagene were downloaded. Leukocytes cell fractions were also downloaded from TCIA for all 755 breast cancer samples. CIBERSORT[17] signature of 22 leukocytes cells was used (LM22).

References

1 Maistro, S. *et al*. Germline mutations in *BRCA1* and *BRCA2* in epithelial ovarian cancer patients in Brazil. *BMC Cancer* **16(1)**, 934, doi: 10.1186/s12885-016-2966-x (2016).

2 Chan, SH. *et al*. Germline Mutations in Cancer Predisposition Genes are Frequent in Sporadic Sarcomas. *Sci Rep*. **7(1)**, 10660, doi: 10.1038/s41598-017-10333-x (2017).

3 Liaw, D. *et al*. Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. *Nat Genet*. **16(1)**, doi: 10.1038/ng0597-64 (1997).

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4 De Queiroz Rossanese, LB. *et al*. APC germline mutations in families with familial adenomatous polyposis. *Oncol Rep*. **30(5)**, 2081-8, doi: 10.3892/or.2013.2681 (2013).

5 Dommering, CJ. *et al*. RB1 mutations and second primary malignancies after hereditary retinoblastoma. *Fam Cancer*. **11(2)**, 225-33, doi: 10.1007/s10689-011-9505-3 (2012).

6 Cetani, F. *et al*. Incidental occurrence of metastatic medullary thyroid carcinoma in a patient with multiple endocrine neoplasia type 1 carrying germline MEN1 and somatic RET mutations. *J Surg Oncol*. **116(8)**, 1197-1199, doi: 10.1002/jso.24793 (2017).

7 Moore, LE. *et al*. Von Hippel-Lindau (VHL) inactivation in sporadic clear cell renal cancer: associations with germline VHL polymorphisms and etiologic risk factors. *PLoS Genet*. **7(10)**, doi: 10.1371/journal.pgen.1002312 (2011).

8 Gray, PN. *et al*. TumorNext-Lynch-MMR: a comprehensive next generation sequencing assay for the detection of germline and somatic mutations in genes associated with mismatch repair deficiency and Lynch syndrome. *Oncotarget*. **9(29)**, 20304-20322, doi: 10.18632/oncotarget.24854 (2017).

9 Milanese, J.S. *et al*. eTumorMetastasis, a network-based algorithm predicts clinical outcomes using whole-exome sequencing data of cancer patients. *bioRxiv* **268680**; doi: https://doi.org/10.1101/268680 (2018).

11 Vanunu, O. *et al*. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol*. **6(1)**, doi: 10.1371/journal.pcbi.1000641 (2010).

12 Hofree, M. *et al*. Network-based stratification of tumor mutations. *Nat Methods* **10(11)**,1108-15, doi: 10.1038/nmeth.2651 (2013).

13 Li, J. et al. Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat Commun* **1**, 34, doi:10.1038/ncomms1033 (2010).

14 Gao, S. *et al*. Identification and Construction of Combinatory Cancer Hallmark-Based Gene Signature Sets to Predict Recurrence and Chemotherapy Benefit in Stage II Colorectal Cancer. *JAMA Oncol* **2**, 37-45, doi:10.1001/jamaoncol.2015.3413 (2016).

15 Huang, D.W. *et al* .Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. **4(1)**, 44-57, doi: 10.1038/nprot.2008.211 (2009).

16 Charoentong, P. *et al*. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep*. **18(1)**, 248-262, doi: 10.1016/j.celrep.2016.12.019 (2017).

17 Newman, M. *et al*. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. **12(5)**, 453–457, doi: 10.1038/nmeth.3337 (2015).

18 Polychemotherapy for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. *Lancet*. **352(9132)**, 930-42, doi: https://doi.org/10.1016/S0140-6736(98)03301-7 (1998).

19 Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*. **452(7187)**, 564-70, doi: 10.1038/nature06915 (2008).

20 Pai, S.G. *et al.* Wnt/beta-catenin pathway: modulating anticancer immune response. J *Hematol Oncol*. **10**, 101, doi: 10.1186/s13045-017-0471-6 (2017).

21 Sarvaria, A. *et al*. B cell regulation in cancer and anti-tumor immunity. *Cell Mol Immunol*. **14(8)**, 662–674, doi: 10.1038/cmi.2017.35 (2017).

22 Yuen, J. G. *et al.* B lymphocytes and cancer: a love-hate relationship. *Trends Cancer*. **2(12)**, 747–757, doi: 10.1016/j.trecan.2016.10.010 (2016).

23 Tsuda, B. *et al*. B-cell populations are expanded in breast cancer patients compared with healthy controls. *Breast Cancer*. **25(3)**, 284–291, doi: 10.1007/s12282-017-0824-6 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

24 Théry, C. *et al*. The cell biology of antigen presentation in dendritic cells. *Curr Opin Immunol*. **13(1)**, 45-51, doi: https://doi.org/10.1016/S0952-7915(00)00180-1 (2001).

25 Gu-Trantien, C. *et al*. CXCL13-producing TFH cells link immune suppression and adaptive memory in human breast cancer. *JCI Insight*. **2(11)**, pii: 91487, doi: 10.1172/jci.insight.91487 (2017).

26 Matkowski, R. *et al*. The prognostic role of tumor-infiltrating CD4 and CD8 T lymphocytes in breast cancer. *Anticancer Res*. **29(7)**, 2445-51 (2009).

27 Hadrup, S. *et al*. Effector CD4 and CD8 T Cells and Their Role in the Tumor Microenvironment. *Cancer Microenviron*. **6(2)**, 123–133, doi: 10.1007/s12307-012-0127-6 (2013).

28 Hanna, R.N. *et al*. Patrolling monocytes control tumor metastasis to the lung. *Science*. **350(6263)**, 985-90, doi: 10.1126/science.aac9407 (2015).

29 Cassetta, L. *et al*. Cancer immunosurveillance: role of patrolling monocytes. *Cell Res*. **26(1)**, 3–4, doi: 10.1038/cr.2015.144.

30 Al Sayed, M.F. *et al*. T-cell-secreted TNF-alpha induces emergency myelopoiesis and myeloid-derived suppressor cell-differentiation in cancer. *Cancer Res*. pii: canres.3026.2018, doi: 10.1158/0008-5472.CAN-17-3026 (2018).

31 Morrow, E.S. *et al*. The role of gamma delta T lymphocytes in breast cancer: a review. *Transl Res*. **203**, 88-96, doi: 10.1016/j.trsl.2018.08.005 (2018).

32 Wu, D. *et al*. Human γδT-cell subsets and their involvement in tumor immunity. *Cell Mol Immunol*. **14(3)**, 245–253, doi: 10.1038/cmi.2016.55 (2017).

33 Aponte-López, A. *et al*. Mast Cell, the Neglected Member of the Tumor Microenvironment: Role in Breast Cancer. *J Immunol Res*. 2584243. doi: 10.1155/2018/2584243 (2018).

34 Cimpean, A.M. *et al*. Mast cells in breast cancer angiogenesis. *Crit Rev Oncol Hematol*. **115**, 23-26, doi: 10.1016/j.critrevonc.2017.04.009 (2017).

35 Pasero, C. *et al*. Highly effective NK cells are associated with good prognosis in patients with metastatic prostate cancer. *Oncotarget*. **6(16)**, 14360-73, doi:10.18632/oncotarget.3965 (2015).

36 Shenouda, M.M. *et al*. Ex vivo expanded natural killer cells from breast cancer patients and healthy donors are highly cytotoxic against breast cancer cell lines and patient-derived tumours. *Breast Cancer Res*. **19(1)**, 76, doi: 10.1186/s13058-017-0867-9 (2017).

37 Marcus, A. *et al*. Recognition of Tumors by the Innate Immune System and Natural Killer Cells. *Adv Immunol* **122**, 91-128, doi:10.1016/B978-0-12-800267-4.00003-1 (2014).

38 Satoh, H. *et al*. Nrf2-deficiency creates a responsive microenvironment for metastasis to the lung. *Carcinogenesis* **31**, 1833-1843, doi:10.1093/carcin/bgq105 (2010).

39 Hurley, P. J. *et al*. Germline Variants in Asporin Vary by Race, Modulate the Tumor Microenvironment, and Are Differentially Associated with Metastatic Prostate Cancer. *Clin Cancer Res*. **22**, 448-458, doi:10.1158/1078-0432.CCR-15-0256 (2016).

40 DePristo, M. A. *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:10.1038/ng.806 (2011).

41 Koboldt, D. C. *et al*. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. **22**, 568-576, doi:10.1101/gr.129684.111 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Competing interests

The authors declare no competing interests.

Author contributions

Manuscript draft: JSM, AN, RM, SD, EW.

Conceptualization and implementation: JSM, CT, EW.

Data acquisition: JSM, JZ, PH, ZM.

Data analysis: JSM, CT.

Data interpretation: JSM, CT, RM, EW.

Table 1 Demographic and clinical characteristics for ER+ breast cancer samples

Variable	Training Set (n=200)		Validation Set 1 TCGA-CPTAC (n=295)		Validation Set 2 TCGA Nature (n=200)	
Clinical Characteristic	**Number of patients**	**%**	**Number of patients**	**%**	**Number of patients**	**%**
Age, years						
Median	59		60		58	
<=59	102	51	149	50.5	105	52.5
>59	98	49	146	49.5	95	47.5
Death						
Yes	29	14.5	33	11.2	26	13
No	171	85.5	262	88.8	174	87
Localization						
Left	102	51	146	49.5	113	56.5
Right	98	49	149	50.5	87	43.5
Stage						
I	37	18.5	53	17.9	30	15
II	108	54	164	55.6	113	56.5
III	40	20	72	24.4	49	24.5
IV	8	4	2	0.7	4	2

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

X	5	2.5	2	0.7	3	1.5
NA	2	1	2	0.7	1	0.5
Subtype						
Luminal A	95	47.5	38	12.9	58	29
Luminal B	42	21	18	6.1	46	23
Unknown	10	5	11	3.7	21	10.5
NA	53	26.5	228	77.3	75	37.5
Relapse						
Yes	30	15	34	11.5	20	10
No	170	85	261	88.5	180	90

Table 2 Prediction accuracy and recall rate for validation sets for breast cancer using the NOG_CSS sets derived from germline mutations

Dataset	Number of samples	Low-Risk		High-risk	
		Accuracy (%)*	Recall (%)†	Accuracy (%)**	Recall (%)††
Training Set	200	93.8	26.5	27.5	36.7
TCGA-Nature	200	94.9	31.1	10.4	65.0
TCGA-CPTAC	295	93.5	38.7	21.0	50.0

Notes:
*Percentage of non-recurred (i.e., non-metastatic) samples in the predicted low-risk group.
†Percentage of the predicted low-risk samples from the non-recurred group.
**Percentage of recurred (i.e., metastatic) samples in the predicted high-risk group.
††Percentage of the predicted high-risk samples from the recurred group.

bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

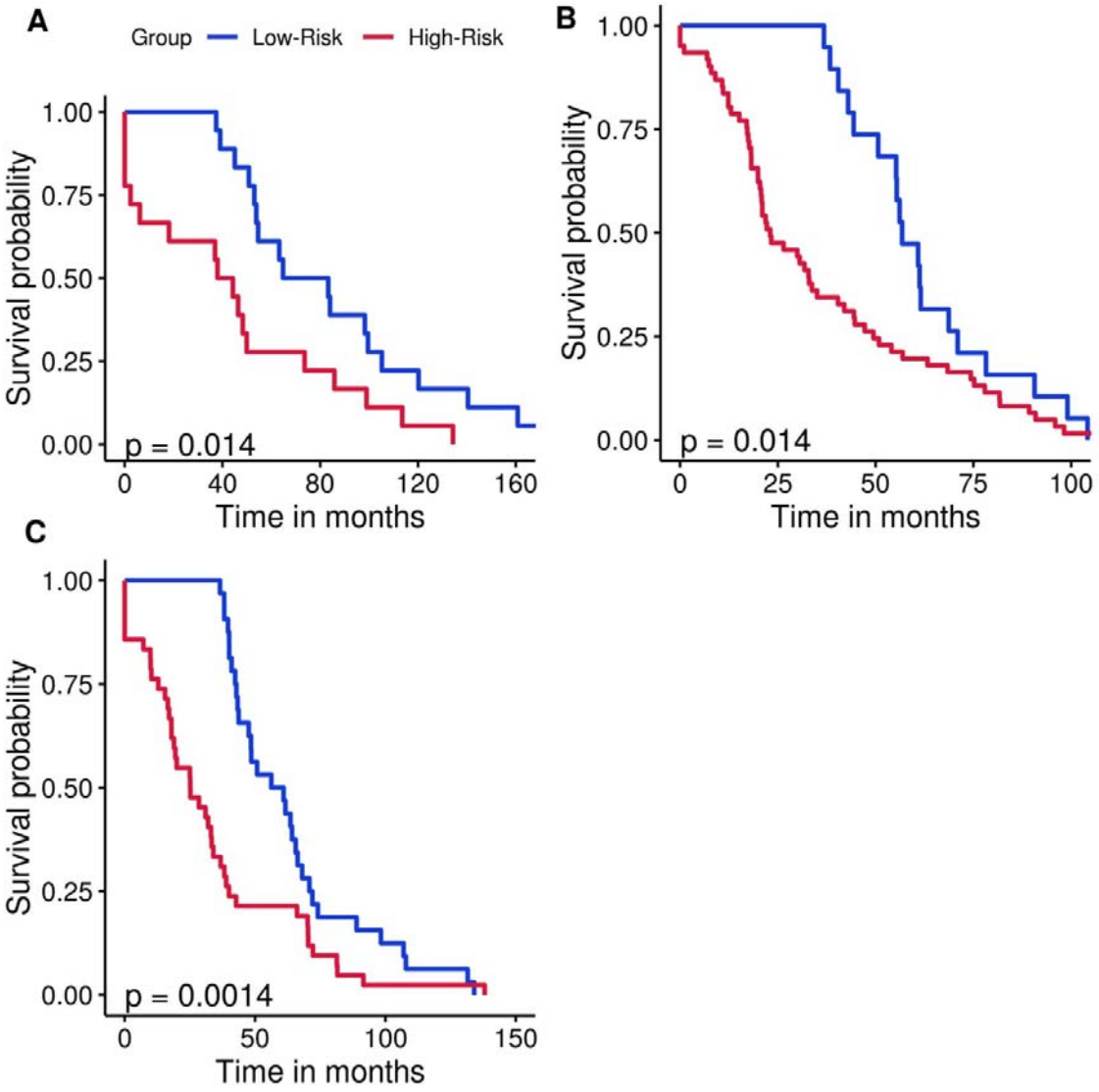
Table 3 Prediction accuracy and recall rate for validation samples for breast cancer using the NOG_CSS sets derived from gene expression of normal tissue

Dataset	Number of samples	Low-Risk		High-risk	
		Accuracy (%)*	Recall (%)†	Accuracy (%)**	Recall (%)††
TCGA-Validation	49	88.9	48.5	66.7	62.5

Notes:
*Percentage of non-recurred (i.e., non-metastatic) samples in the predicted low-risk group.
†Percentage of the predicted low-risk samples from the non-recurred group.
**Percentage of recurred (i.e., metastatic) samples in the predicted high-risk group.
††Percentage of the predicted high-risk samples from the recurred group.

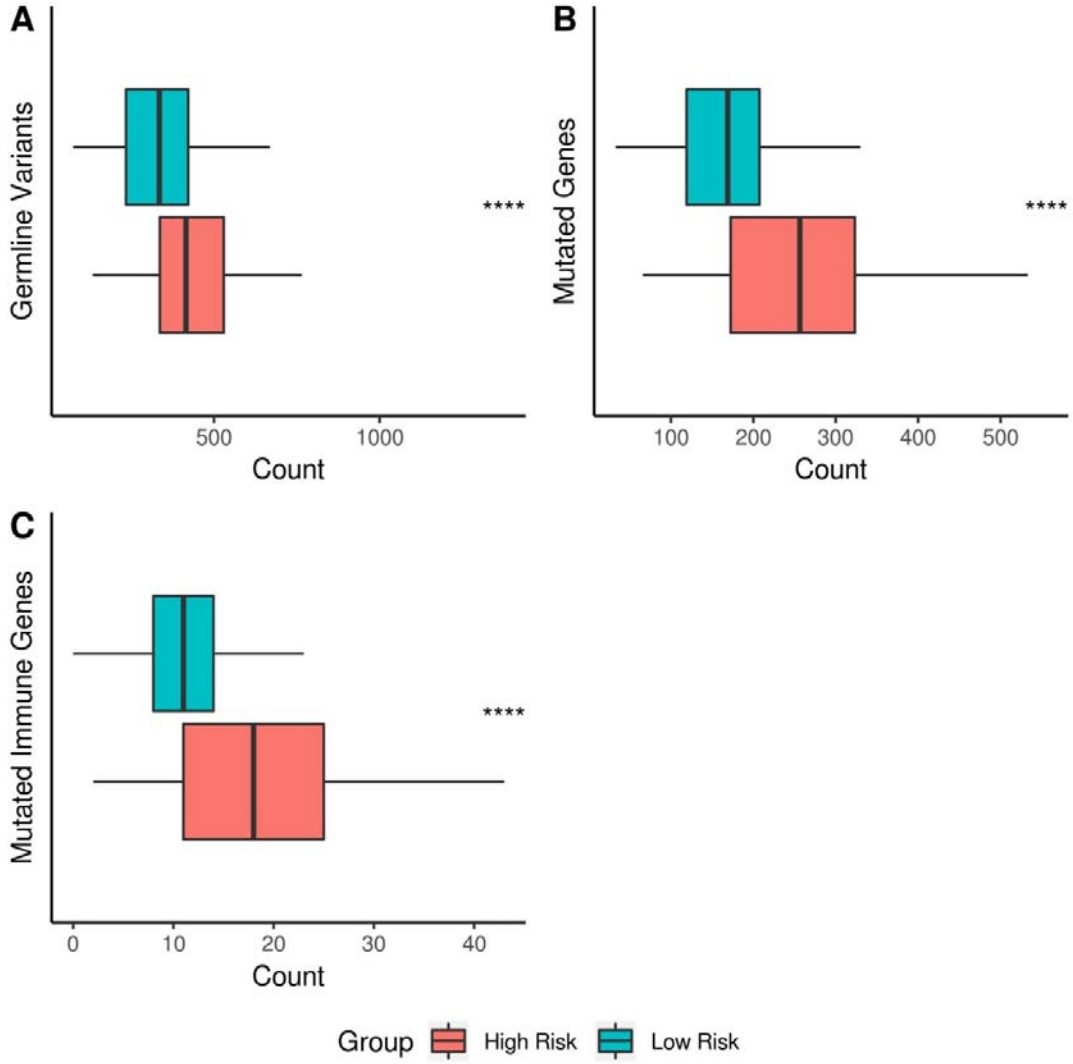
bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1 Kaplan–Meier curves of the risk groups for breast cancer patients with 10-year disease-free survival predicted by the NOG_CSS sets. Samples without DFS time or who couldn't be predicted were removed. NOG_CSS sets derived from germline mutations in (**A**) the training set, (**B**) the validation set, TCGA-Nature and (**C**) the validation set, TCGA-CPTAC. Blue and red curves represent low- and high-risk groups, respectively. P-values were obtained from the χ^2-test.



bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

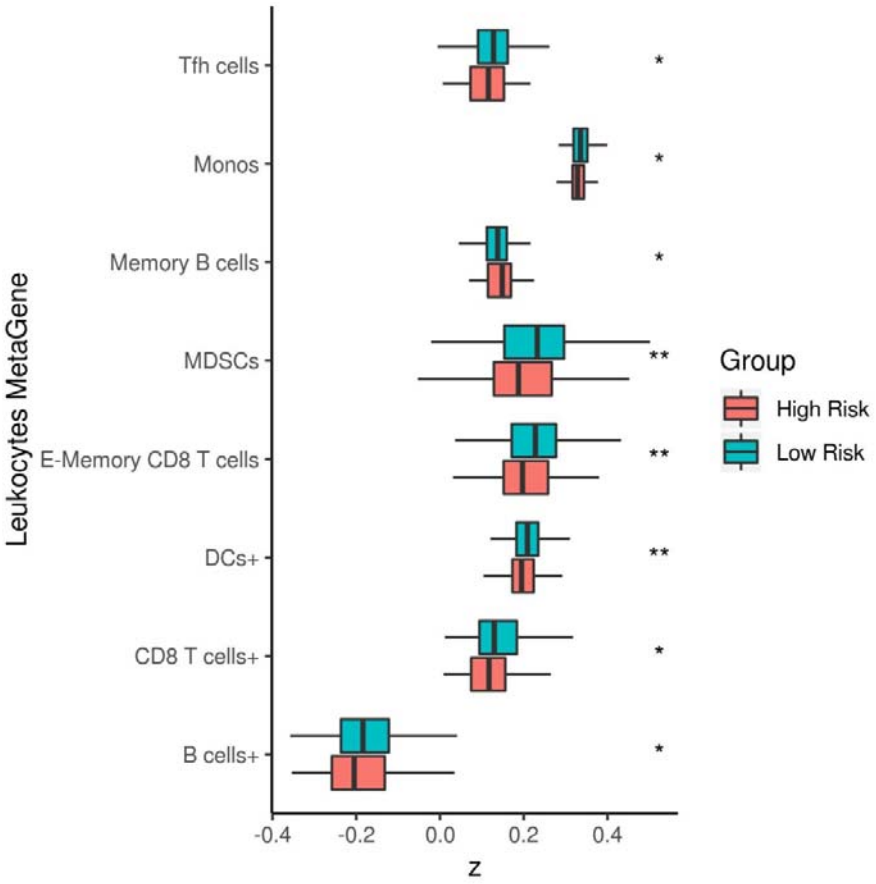
Figure 2 Boxplots comparison of functional germline mutated variants and genes for predicted risk groups. Samples who couldn't be predicted were removed. (A) Functional germline variants **(B)** Functionally mutated genes **(C)** Germline mutated immune genes. P-Values were obtained from Student's t-test. P-Value significance: **** < 0.0001.



bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3 Boxplots comparison of leukocytes metagenes and cell fractions for predicted risk groups. Samples who couldn't be predicted were removed. For a complete analysis, see Supplementary Figure 1. (**A**) Leukocytes gene expression (**B**) Leukocytes cell fraction. P-Values were obtained from Student's t-test. P-Value significance: * < 0.05, ** < 0.01.

Fig 3A



bioRxiv preprint doi: https://doi.org/10.1101/312355; this version posted January 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

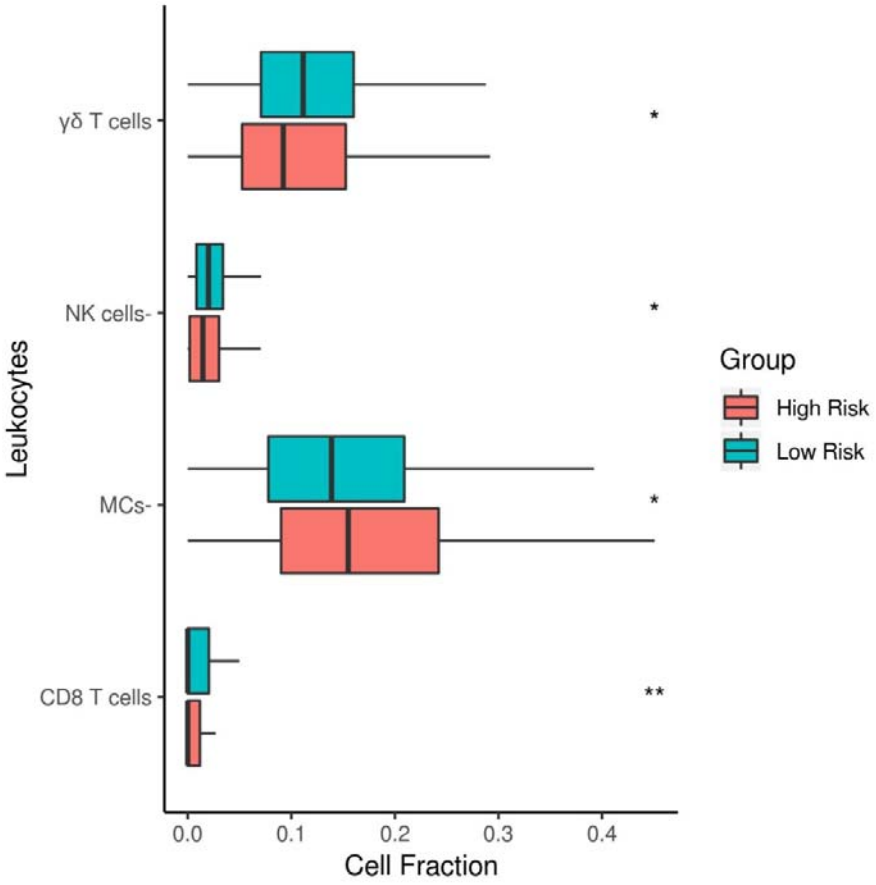



Fig 3B