Germline genomes of breast and lung cancer patients significantly predict clinical outcomes

Jean-Sébastien Milanese[1,2], Chabane Tibiche[1], Naif Zaman[1], Jinfeng Zou[1], Pengyong Han[3], Zhigang Meng[3,5], Andre Nantel[1], Arnaud Droit[2], and Edwin Wang[1,3,4,6*]

1. National Research Council Canada, 6100 Royalmount Ave, Montreal, Canada, H4P 2R2
2. Genomics Center, Centre Hospitalier Universitaire de Québec - Université Laval Research Center, Quebec, Canada, G1V 4G2
3. Department of Biochemistry & Molecular Biology, Medical Genetics, and Oncology, University of Calgary, 3330 Hospital Dr. NW, Calgary, Canada, T2N 4N1
4. Alberta Children's Hospital Research Institute and Arnie Charbonneau Cancer Research Institute, University of Calgary, 3330 Hospital Dr. NW, Calgary, Canada, T2N 4N1
5. Chinese Academy of Agricultural Science, No. 12 Zhongguangcun South St., Haidian District, Beijing, 100086, China
6. Department of Medicine, McGill University, 3605 Mountain St, Montreal, Canada H3G 2M1

*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

Abstract

Germline genetic variants such as BRCA1/2 could play an important role in tumorigenesis and even clinical outcomes of cancer patients. These germline mutations have been used for genetic counselling and testing to establish as a standard-of-care procedure in the management of cancer patients. However, only a small fraction (i.e., 5-10%) of cancer patients whose clinical outcomes have been associated with inherited mutations (e.g., BRCA1/2, APC, TP53, PTEN and so on). It is challenging to use all the inherited mutations to predict the clinical outcomes for the majority of cancer patients. To deal with this issue, we applied our recently developed algorithm which enables to construct predictive models using genome sequencing data to ER+ breast (n=755) and lung (n=436) cancer patients. Gene signatures derived from functionally mutated genes in germline genomes significantly distinguished recurred and non-recurred patients in two ER+ breast cancer independent cohorts (n=200 and 295, $P=1.0x10^{-7}$) and the lung cancer cohort (n=176, $P=1.8x10^{-2}$), and outperformed predictions using clinical factors. The signature genes are predominately impairing patients' immune functions such as T cell function, antigen presentation and cytokine interactions (i.e., recurred patients have more such mutations). These results suggest that inherited mutations weakening patients' immune system could impact tumorigenesis and evolution in not only genetically- (i.e., breast) but also environmentally-driven (i.e., lung) cancers. Most importantly, germline genomic information could be used for developing non-invasive genomic tests for predicting patients' outcomes (or drug response) in breast and lung cancers, and even other cancer types and complex diseases.

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) thus initiates tumor progression and recurrence. In general, cancer recurrence and metastasis are the result of the interactions of multiple mutated genes. We hypothesized that mutagenic processes are essentially blind or non-purposeful, however, to drive cancer progression or metastasis, new mutations will be selected if they could integrate into the pre-existing genomic landscape (i.e., germline mutations or germline genetic variants) to trigger or activate a cancer process. This means that pre-existing germline genetic variants could provide a profound constraint on the evolution of tumor founding clones and subclones, and therefore, have a contingent effect on tumor evolution and even patient outcomes. Family history remains one of the major risk factors that contribute to cancer and further, 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 of them 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*, whose mutation results in Cowden syndrome, *APC*, which is linked to familial adenomatous polyposis and the Retinoblastoma gene *RB1*. Two distinct types of multiple endocrine neoplasias are associated with the *RET* and *MEN1* genes while *VHL* alterations result in kidney and other types of cancer. Finally, Lynch syndrome, a form of colorectal cancer, is linked to *MSH2*, *MLH1*, *MSH6*, *PMS2*, and *EPCAM*. Genetic tests based on these highly-penetrant gene mutations have shown their usefulness, but they can explain only a small fraction (5-10%) of the 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 has been unknown that to what extent the germline genomes affect tumorigenesis, tumor evolution and even patients' clinical outcome. We have previously shown that tumor founding clone mutations are able to predict tumor recurrence[3]. Here, we reasoned that the collective impact of germline genetic variants/mutations in cancer patients might largely determine tumorigenesis, evolution and even patients' clinical outcomes. As tumor heterogeneity becomes a more prominent concept, it is our belief that germline genetic mutations act in a combination with somatic mutations to modulate tumorigenesis and

metastasis. Each patient's germline genetic mutations' combination predisposes specific biological/signaling pathways (even phenotypes) that would lead to diverse clinical outcomes of cancer patients. Therefore, the germline genomic landscape of cancer patients could be used as a predictive tool in order to inform clinicians as to when and how the disease might progress. Thus, germline genetic mutations could offer a new non-invasive genetic testing approach because they can be determined using a blood or saliva sample. However, thus far, clinical outcome predictions using cancer germline genomic information have not been demonstrated in many cancer types. The increasing availability of genome sequencing data provide opportunities to develop predictive models that can interpret and translate these complex genomic alterations into clinical outcomes.

In this study, we showed that the collective germline genetic mutations of breast and lung cancer patients enable to predict tumor recurrence by applying a recently developed method, eTumorMetastasis[3], to 755 breast and 436 lung cancer patients, respectively. Further, the germline mutated variants which were associated with tumor recurrence in both cancers could impair the adaptive immune functions of cancer patients. These results highlight the important role of germline genetic mutations in tumor evolution and recurrences.

Results

Germline genetic variants predicted breast cancer recurrence

To exam if germline genetic mutations are able to predict tumor recurrence, we used the whole-exome sequencing data (i.e., from the NCI Genomic Data Commons, GDC) of the healthy tissues from 755 ER+ breast patients by applying our recently developed method, eTumorMetastasis[3]. ER+ subtype represents ~70% of the breast cancer patients, thus, in this study, we used ER+ subtype patients for this cohort. The GDC data portal contains a small number of tumor samples for other subtypes, thus, we did not use them in this study. The demographic table of the breast cancer cohort is represented in Table 1.

We hypothesized that pre-existing germline genetic variants of a cancer patient could complementary work together with somatic mutations (i.e., somatic mutations are evolutionary selected to work with the pre-existing germline genetic variants) to initiate tumorigenesis and metastasis. This is the underlying concept of eTumorMetastasis, and also let us to hypothesize that the pre-existing germline genetic variants of cancer patients could

have predictive power for metastasis and clinical outcomes. The eTumorMetastasis contains 3 components (1) using a network-based approach[4,5] to smooth the data on a cancer type specific signaling network, data are the functionally mutated genes; (2) using our previously developed method, MSS (Multiple Survival Screening)[6] for identifying biomarkers and (3) using our previously developed method for better predictions via combining biomarkers[7]. The detailed procedure of the eTumorMetastasis and network construction were described previously[3]. To apply the eTumorMetastasis, briefly we first annotated the functional mutations 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 metastasis (see Methods), and then mapped the functionally germline-mutated genes on the metastasis network mentioned above. Finally, gene signatures (i.e., biomarkers) were obtained using 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 gene signatures), which could distinguish recurred and non-recurred breast tumors. By applying eTumorMetastasis to the germline genomes of the 200 patients, we identified 18 NOG gene signatures (Supplementary Tables 1 and 2) for ER+ breast cancer. Each of them contains 30 genes repenting a cancer hallmark such as apoptosis, cell proliferation, metastasis, and so on. We previously showed that multiple gene signatures representing distinct cancer hallmarks could be identified from one training cohort[6], furthermore, ensemble-based prediction using multiple gene signatures representing distinct cancer hallmarks enabled to significantly improve prediction performance[7]. Thus, we used the 18 NOG gene signatures to construct a NOG_CSS set (i.e., NOG-based Combinatory Signature Set) using a testing set of 60 samples based the method we developed previously[7]. We then used the NOG_CSS set 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 set significantly distinguished recurred and non-recurred breast tumors in two validations sets: 200 (ER+ Nature-Set, $P=3.4x10^{-2}$), 295 (ER+ TCGA-CPTAC independent set, $P=1.0x10^{-7}$). These results suggested that germline genetic mutations are significantly correlated with tumor metastasis and supported our hypothesis that the original germline genomic landscape of a cancer patient has a significantly impact on clinical outcome of its tumor. Importantly, sequencing a patient's blood or saliva sample could provide a convenient, timely and

noninvasive way to predict patient outcome in a clinical environment.

To compare the prediction performance of the NOG_CSS set with clinical factors, we conducted relapse-free survival analysis of clinical factors using the 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 that derived from the germline NOG_CSS set ($P=1.0x10^{-7}$). These results suggest that gene signatures derived from the germline genomic information have a better predictive performance than clinical factors.

Germline genetic variants predicted lung cancer recurrence

We demonstrated that germline genomic information significantly predicted the prognosis of breast cancer patients. Breast cancer represents a genetically-driven cancer type, we further asked if the germline genomic information of cancer patients was able to predict the clinical outcomes of an environmentally-driven cancers such as lung cancer. Lung cancer is a well-known cancer type which is caused by tobacco smoking (e.g., more than 80% of the lung cancer patients are smokers).

To exam if germline genetic mutations are able to predict lung tumor recurrence, we used the whole-exome sequencing data from the GDC data portal of the healthy tissues from 436 lung cancer patients. Because adenocarcinoma, a subtype of lung cancer, has a larger number of samples in the GDC, we used only adenocarcinoma in this study. Table 3 shows the demographic table of this cohort. The same method used in breast cancer were extended to lung cancer samples except that we constructed a lung cancer-specific molecular network for metastasis (see Methods and Supplementary Methods).

By applying eTumorMetastasis to the germline genomes of 200 lung cancer patients, we identified 14 NOG gene signatures (Supplementary Tables 5 and 6). Furthermore, we used the 14 NOG gene signatures to construct a NOG_CSS set using the testing set containing 60 samples. As shown in Fig 2 and Table 4, the germline-derived NOG_CSS set significantly distinguished recurred and non-recurred lung tumors in the validation set containing 176 samples ($P=1.8x10^{-2}$). These results suggested that germline genetic mutations enables to predict prognosis in lung cancer as well. Furthermore, the pre-existing germline genomic landscapes of cancer patients could play an important role in shaping the tumor metastasis regardless of the cancer types which are either genetically- or environmentally-driven. It has

been known that only 10-15% of the heavy smokers develop lung cancer in their whole life[8-9], implying that germline genome plays an important role in tumorigenesis for smokers. Our results suggest that germline genome also plays an important role in metastasis for lung cancer.

To compare the prediction performance of the NOG_CSS set with clinical factors, we conducted relapse-free survival analysis of clinical factors using the Cox proportional hazards regression model. The best p-value (i.e., $P=7.0x10^{-2}$ log-rank test) using covariate models (Supplementary Table 7) was not better than that derived from the germline CSS set ($P=1.8x10^{-2}$). These results suggest that gene signatures derived from the germline genomic information have a better predictive performance than clinical factors in lung cancer.

Predictive germline genetic variants could impair the adaptive immune system of breast and lung cancer patients

To further understand why germline genomic landscapes of cancer patients are predictive for tumor recurrence, we ran enrichment analyses for the genes of the NOG signatures of breast and lung cancers, respectively, using DAVID[10]. Interestingly, for both cancer types, most of the enriched biological pathways and Gene Ontology (GO) terms have two categories: immune system and cell division (Supplementary Tables 9 and 10), but most of the pathways and GO terms (i.e., more affected in the recurred group) are associated with the adaptive immune system. For example, antigen processing and presentation, cytokine-cytokine receptor interaction, T cells co-stimulation are well known immune related process while cell division and cell cycle are related to cell division process. T cell helps eliminating cancer cells by recognizing and binding to an antigen(s) or neoantigens which have arisen as a result of somatic mutations in cancer cells but not normal cells. To be seen by T cells, neoantigens muse be presented on cancer cell surface by the antigen processing and presentation system. Cytokine and T cells co-stimulation regulate the functions of T cells in the adaptive immune system, while Wnt signaling pathway is associated with tumor immune cell infiltration. Therefore, functionally mutations or deregulation of either of these pathways or biological processes mentioned above could affect the results of T cell attacking to cancer cells and then tumor progression and recurrence. These results suggest that the germline genomes of cancer patients encode inherited variants which could dysregulate immune system to impair the adaptive immune system in certain degrees so that the inherited variants could influence tumor progression, metastasis and patient outcomes.

Discussion

Thus far, this is the first study which has shown that the germline genomes of cancer patients predicted clinical outcomes. These results suggest that the germline genetic variants play an important role in shaping tumor evolution and recurrence. The enrichment analysis of the NOG_gene signatures derived from cancer patients' germlines suggested that germline genomes of the recurred patients encode activated signaling pathways which are associated with inflammation and cell adhesion which could promote tumor recurrence.

Our understanding of the fundamental cancer biology of metastasis is limited. Several reports have provided evidence that germline mutations of cancer patients modulate metastasis through tumor microenvironments. For example, it is known that germline-encoded receptor variants could trigger innate immune response in cancer patients[11]. Lung cancer patients with a germline mutation in Nrf2 have a good prognosis due to the fact that Nrf2 exerts anti-metastatic activity by regulating the inflammatory status and redox balance of the hematopoietic and immune systems of cancer patients[12]. In prostate cancer, patients with a germline variant of the ASPN D locus are highly associated with poorer oncologic outcomes[13], probably through its interactions with TGF-β. These studies highlight the impacts of cancer patients' genetic makeup on metastasis and provide rationales for further studies on the impact of germline genomic landscapes on clinical outcomes of carcinogenesis.

Our study further suggests that metastasis is predictive based on the germline genomic landscapes of cancer patients. We found that germline mutations related to cell division, immune cell infiltration and T cell activities are predominately predictive for tumor recurrence. For example, more mutations in the antigen processing and presentation pathway could impair the presentation of neoantigens of cancer cells so that T cells can't recognize the tumor cells allowing them to escape immune systems' monitoring. Mutations in cell division process could introduce more mutations during cell division. Activation of Wnt pathway can block the infiltration of immune cells within tumors[14]. These results suggest that the germline genome of cancer patients have pre-existing mutations which could have impact on immune system, cell division and immune microenvironment and then affect metastasis and patient outcome. Taken together, the germline genomic landscape of a cancer patient might provide not only a substantial constraint on tumor recurrence for tumor cells, but also constrains for

tumor microenvironments, which ultimately has a profound impact on patient outcomes. Furthermore, we proposed that the future study of germline genomics in cancer biology could be important. Traditionally, germline genomics have been largely ignored in the cancer genomic community; for example, most of the cancer genomic studies including the GDC have often focused on somatic mutations only while germline mutations have been filtered out before formal analysis of tumor genome sequencing data.

The demonstration that germline exome sequencing data can predict cancer patients' outcomes suggests that we could develop non-invasive genomic tests using the blood or saliva samples of cancer patients to determine cancer prognosis for guiding clinicians in making treatment decisions. It is also possible to develop non-invasive genomic tests for drug response (i.e, germline genomic data could predict anti-CTLA4 response to melanoma patients, unpublished data). 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 forecasting of cancer recurrences in a quick, convenient and minimally invasive manner. We showed that genome sequencing of patient germlines might provide efficient, non-painful and convenient genomic test for predicting tumor recurrence. Thus, germline genomic information could be used for genetic counselling and testing to establish as a standard-of-care procedure in the management of cancer patients.

Methods

We obtained whole-exome sequencing data of the germlines for breast and lung 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). For lung cancer, a training set of 200 samples, a testing set of 60 samples and one validation set of 176 samples (Supplementary Table 8). Raw sequence reads from healthy samples of cancer patients were processed using the methods described previously[3]. Variant calling was then performed using Varscan2[15].

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

To identify NOG_gene signatures using the functional mutated genes of breast cancer patients' germline genomes, we followed the eTumorMetastasis[3] method. Briefly, we constructed a breast-specific metastasis network based on the methods described previously[3]. For each patient, we used its germline functionally mutated genes as seeds on the breast cancer-specific metastasis network to perform network propagation and then identify NOG_gene signatures.

For lung cancer, to identify NOG_gene signatures, we applied the same methodology which was used in breast cancer. Briefly, we constructed a lung cancer-specific metastasis network based on the methods described previously[3]. (see Supplementary Methods for details). Once again, for each patient, we used its germline functionally mutated genes as seeds on the lung-specific metastasis network to perform network propagation and then identify NOG_gene signatures.

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 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).

4 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).

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

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

7 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).

8 Crispo, A. *et al.* The cumulative risk of lung cancer among current, ex- and never-smokers in European men. *Br J Cancer.* **91(7)**, 1280-6, doi:10.1038/sj.bjc.6602078 (2004).

9 Brennan, P. *et al.* High cumulative risk of lung cancer death among smokers and nonsmokers in Central and Eastern Europe. *Am J Epidemiol.* **164(12)**, 1233-41, doi:10.1093/aje/kwj340 (2006).

10 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).

11 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).

12 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).

13 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).

14 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).

15 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).

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
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 (%)††
TCGA-Nature	200	91.4	53.3	12.7	40.0
TCGA-CPTAC	295	97.4	71.3	35.4	67.7

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.

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. 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. Kaplan–Meier curves have not been made for samples which do not have follow-up time. Blue and red curves represent low- and high-risk groups, respectively. P-values were obtained from the χ^2-test.

Fig 1A

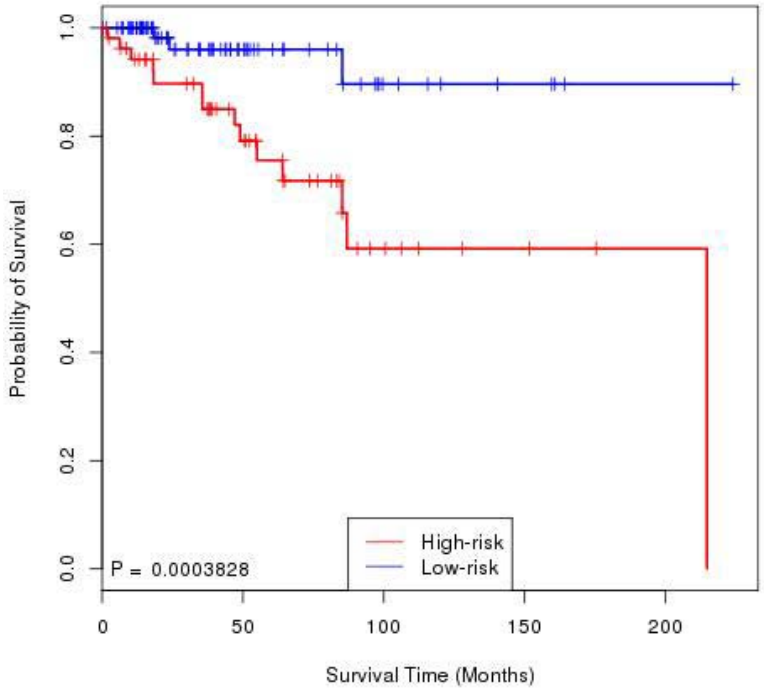


Fig 1B

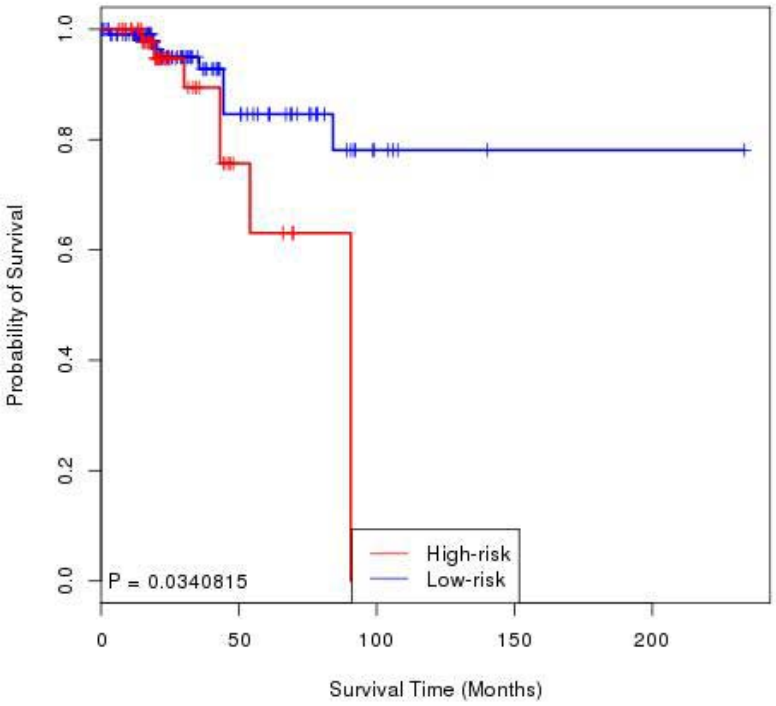

Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
P = 0.0340815
High-risk
Low-risk
0
50
100
150
200
Survival Time (Months)

Fig 1C

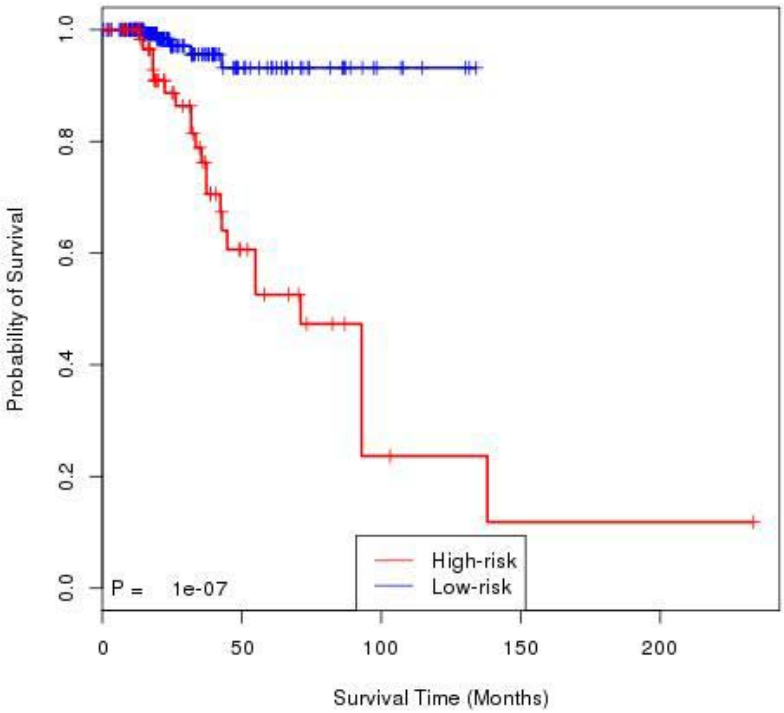

Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
P = 1e-07
High-risk
Low-risk
0
50
100
150
200
Survival Time (Months)

Table 3 Demographic and clinical characteristics for lung cancer samples

Variable	Training Set (n=200)		Validation Set (n=176)	
Clinical Characteristic	**Number of patients**	**%**	**Number of patients**	**%**
Age, years				
Median	65		68	
<=65	102	51	75	42.6
>65	90	45	97	55.1
NA	8	4	4	2.3
Gender				
Male	94	47	76	43.2
Female	106	53	100	56.8
Death				
Yes	65	32.5	67	38.1
No	135	67.5	109	61.9
Localization				
Left	74	37	77	43.8
Right	117	58.5	97	55.1
NA	9	4.5	2	1.1
Stage				
I	103	51.5	100	56.8
II	46	23	41	23.3
III	38	19	23	13.1
IV	7	3.5	10	5.7
NA	6	3	2	1.1
Relapse				
Yes	100	50	91	51.7
No	100	50	85	48.3

Table 4 Prediction accuracy and recall rate for validation sets for lung cancer using the NOG_CSS sets derived from germline mutations

Dataset	Number of samples	Low-Risk		High-risk	
		Accuracy (%)*	Recall (%)†	Accuracy (%)**	Recall (%)††
Validation Set	176	73.4	36.4	75.0	15.6

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.

Figure 2 Kaplan–Meier curves of the risk groups for lung cancer patients predicted by the NOG_CSS sets. NOG_CSS sets derived from germline mutations in (A) the training set, (B) the validation set. Kaplan–Meier curves have been not made for samples which do not have follow-up time. Blue and red curves represent low- and high-risk groups, respectively. P-values were obtained from the χ^2-test.

Fig 2A

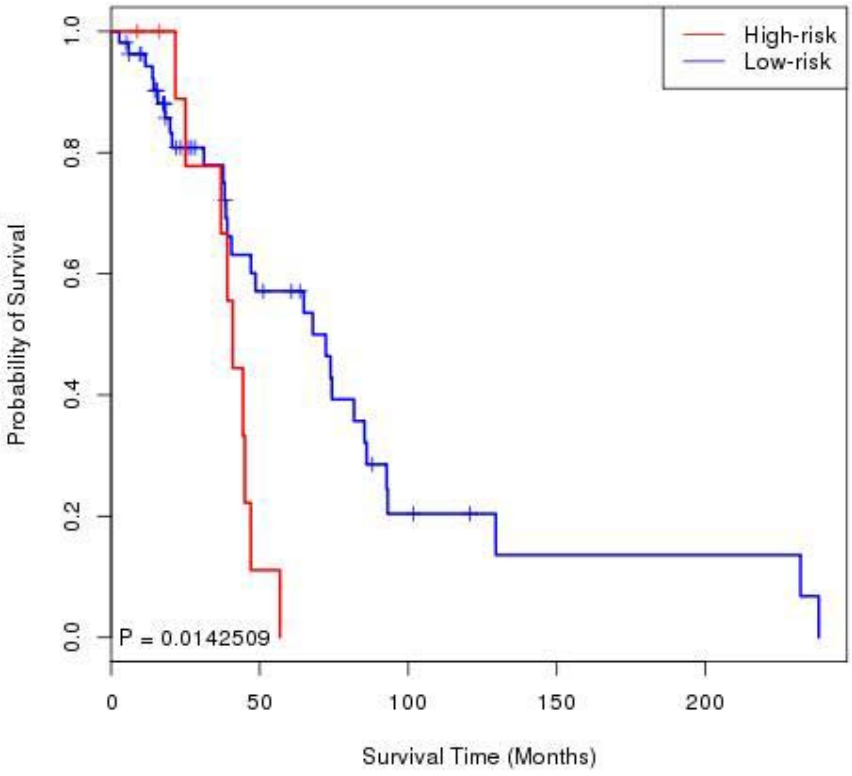
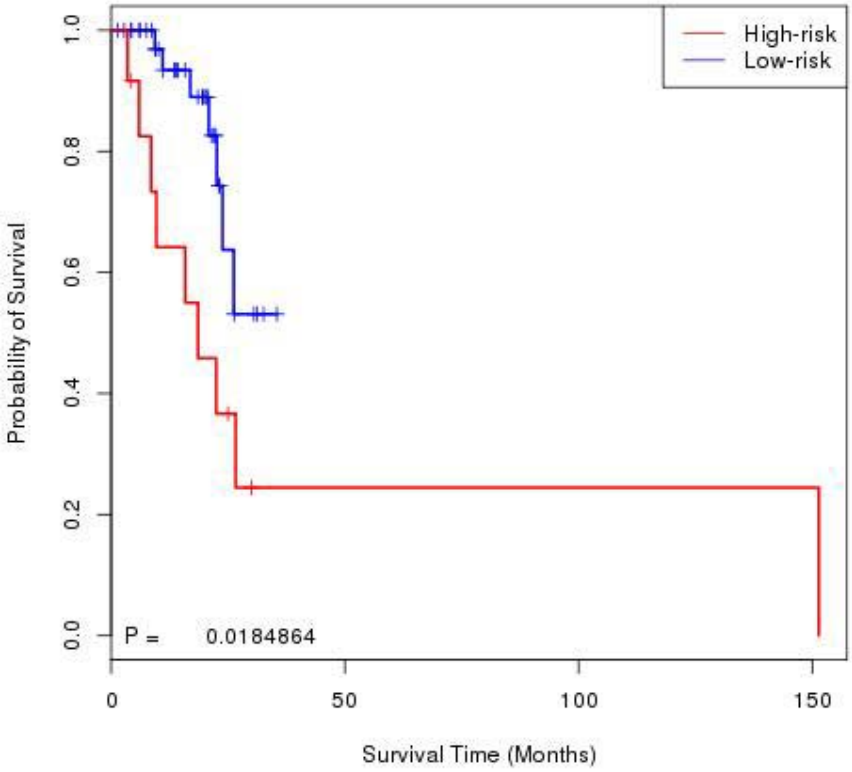


Fig 2B


High-risk
Low-risk
Probability of Survival
1.0
0.8
0.6
0.4
0.2
0.0
P = 0.0184864
0
50
100
150
Survival Time (Months)