Research Plan

Summary

Inflammatory breast cancer (IBC) is a rare, aggressive form of breast cancer that is characterized by a highly invasive and metastatic phenotype. It accounts for approximately 5% of newly diagnosed breast cancers. The diagnosis is based on clinical features including rapid development of skin erythema and edema due to cancer cells invading the lymphatic vasculature of the skin. Numerous attempts, including our previous research, have failed to identify gene expression or DNA copy number alterations that distinguish IBC from non-IBC [1,2]. Currently, there is no molecular diagnostic test for IBC and the molecular mechanisms that underlie the highly invasive and metastatic phenotype remain unknown. However, DNA sequence level alterations in IBC have not been examined systematically and no exome or whole genome sequence data exists for this disease.

We hypothesize that DNA sequence changes in the coding or non-coding regions of the genome may be responsible for the unique phenotype of IBC. The goal of this project is to perform whole genome sequencing of IBC specimens and compare these to non-IBC in order to identify IBC-specific sequence alterations and validate candidate genomic markers on an independent cohort of IBC cases. The specific aims of this 1-year project are:

Aim 1 is to evaluate the genomic landscape of IBC by whole genome sequencing of 20 IBC and matched normal samples performed through JAX under the supervision of Dr Chuang, to identify germ-line and somatic variants (SNPs, indels and structural alterations). We will search for alterations in both expressed genes and noncoding sequences as potential drivers. Noncoding annotations will be derived from ENCODE data and we will use FunSeq developed by Dr Gerstein's lab to categorize both coding and noncoding variants as potential drivers.

Aim 2 is to compare the genomic landscape of IBC with non-IBC. We will perform comparative analysis using TCGA and ICGC breast cancer sequence data (n=216 whole genome, n>600 whole exome) adjusted for by molecular subtypes. Drs Ueno and Gerstein are also members of the ICGC group as tissue provider and analysis sites, respectively. Variants will be mapped to protein-protein interaction and regulatory networks and pathway analysis will be performed to assess if biological processes, rather than recurrent single anomalies, could differentiate IBC from non-IBC. Dr Pusztai's group has entered into a collaborative research agreement with IBM Watson Genomics to apply Watson artificial intelligence software tool to interpret the multiplicity of genomic anomalies in a cancer and draw therapeutic and biological hypothesis. We will also explore the possibility that a genome-wide metric such as overall mutation load, genome entropy or Mutant Allele Tumor Heterogeneity (MATH) define the difference between IBC and non-IBC.

Aim 3 is validation of the candidate markers in independent IBC and non-IBC samples. We have received approval from SWOG and the NCI to perform next generation sequencing on the samples of the SWOG S0800 neoadjuvant clinical (n=200) trial that includes n=60 IBC. RNA and DNA extraction has started from these specimens which will be used to validate candidate markers.

This proposal will be the first whole genome sequencing effort of IBC and will bring together clinical breast cancer experts, bioinformaticians and geneticists from 3 different SWOG institutions (MDACC, Yale and JAX) and leverage a SWOG tissue resource. The S0800 is the only clinical trial that included a large number of IBC cases and therefore represents a unique resource. This project will also create a new genomic data resource for the broader research community. The results could lead to new objective molecular diagnostic criteria for IBC, similar to E-cadherin loss that defines invasive lobular carcinomas of the breast, and minimize the subjectivity of the diagnosis. The findings will also advance our understanding the genomic underpinnings of cancer metastasis and invasion and ultimately may lead to the development of novel therapeutic strategies.

Research Plan

Tissues: All human tissues that are necessary to complete this study are currently available in the laboratory of Dr Pusztai at Yale Cancer Center. Genomic DNA from 20 snap frozen, newly diagnosed IBC and matching normal tissues were received from the Morgan Welch Inflammatory Breast Cancer Research Program and Clinic at MDACC to discover IBC-specific sequence alterations. These specimens were individually reviewed by a breast pathologist and our clinical investigator Dr Ueno for accuracy of diagnosis and tumor cellularity > 60%. The samples were collected under an IRB approved study and were transferred to Yale under an material transfer and research agreement. For validation, we will use formalin fixed paraffin embedded cancer tissues that were collected before any therapy during the S0800 randomized neoadjuvant clinical trial. Dr Pusztai is the correlative science PI of the study and both SWOG and the NCI has approved the use of these tissues for molecular characterization under study title: "*Whole Exome Sequencing of DNA from Pre-Chemotherapy Needle Biopsies of Triple Negative and Inflammatory Breast Cancers Enrolled on the S0800 Trial.*" This tissue analysis project will be performed under the Yale IRB protocol: *Elucidation of Breast Cancer Biology Using Next Generation Sequencing (HIC #1406014226)*.

Preliminary Results: One microgram germline and tumor DNA were used from 2 discovery cases to perform 150 bp, paired-end whole genome sequencing with median coverage of 60X using the Illumina HiSeq 2500 sequencing platform at The Jackson Laboratory for Genome Medicine (JAX) in Farmington, CT to demonstrate feasibility. We identified germline and somatic single nucleotide variants, INDELs and large structural alterations in these two samples. Furthermore, we prioritized somatic SNVs and investigated their genomic annotations by using our variant prioritization tool FunSeq2. In addition, we also prioritized and compared the genomic annotations of variant present in non-BC samples. These variant sets were curated from previously published result [3]. This comparison study suggest that SNVs in IBC influence genes, which are common across all non-IBC samples as well as certain specific genes affected only in IBC samples (Figure1). We also observed large number of germline SVs and relatively small number of somatic SVs in both IBC samples (Table1). However, somatic SVs were comparatively larger in size than germline SVs.

Analysis Plan:
Aim 1*. Identification of variants:* As part of the International Pan-Cancer Analysis of Whole Genomes (PCAWG) Initiative, the Gerstein Lab is involved in improving and standardizing variant calling methods, developing new methods for SV calling, identifying noncoding drivers and network and pathway analysis. Tools developed for this initiative will be applied to the current project. We will identify germ-line single nucleotide variants (SNVs) and INDELs using GATK haplotype caller [4]. Furthermore, we will utilize MuTect [5] and Strelka [6] to obtain highly confident somatic SNVs and Indels, respectively. In addition, we will run CREST to call germline and somatic structural variations [7]. ***Systematic annotation of the variants:*** We will annotate variants in the coding and non-coding regions using FunSeq2 [8]. Noncoding annotations derived primarily from ENCODE will include transcription-factor binding sites, DNA-hypersensitive sites, chromatin marks by histone binding, predicted enhancer regions, miRNA and pseudogenes [9]. **Identification of c*andidate coding and noncoding drivers***: We developed FunSeq to prioritize both coding and noncoding variants that could be potential drivers [10]. Briefly, the tool identifies potential regions of high functional impact in noncoding regions by understanding patterns of natural variation in human genomes and comparing these patterns in disease cases. We identified regions in the genome under purifying selection that are enriched for rare alleles using variation data from 1092 individuals (Phase1 of 1000 genomes project [11]). Such regions that we dubbed sensitive and ultrasensitive

regions highlight regions that are under strong constraint. Mutations identified in such functionally important regions will be considered potential cancer drivers. We have used this method to successfully identify potential noncoding driver mutations in prostate cancer genomes [10]. We have now enhanced this pipeline to include expanded enhancer predictions, gain and loss of function motif analyses for TF-binding and identification of genes that may be affected by such regulatory variants [8]. We will identify candidate coding driver mutations by integrating analyses based on our previously developed pipelines specifically for coding region of genome. We will run Variant Annotation Tool (VAT), which provides transcript-specific annotations and annotates mutations as synonymous, missense, nonsense or splice-site disrupting changes [12]. We will further prioritize loss-of-function (LOF) mutations by employing our ALOFT pipeline. In addition, will also utilize our computation method - NetSNP [13], which incorporates multiple network and evolutionary properties to quantify indispensability of each gene in the genome.

Aim 2. *Comparison with non-IBC:* We will use breast cancer sequencing data from the TCGA [14] (n>600 whole exomes) and ICGC [15] including (n=216 whole genomes), to identify candidate sequence-based markers that distinguishes IBC from non-IBC. As part of the PCAWG initiative, we have obtained permission and already have access to all pubic and controlled access data from the TCGA and ICGC consortiums. All IBC non-IBC comparisons will be balanced for molecular subtype distribution. We will perform signature analysis to distinguish molecular signature pattern prevalent among various subtypes of breast cancers. We will also employ our coding and non-coding prioritization scheme to identify common and distinct driver events between IBC and non-IBC samples. In addition, we will compare germline and somatic SV profiles (type of SVs and their overlap with genomic elements) and their underlying mechanism to distinguish between IBC and non-IBC samples. ***Network and pathway analysis:*** We will use the regulatory element-target gene pairs to connect non-coding variants into a variety of networks -- e.g. regulatory network, metabolic pathways, etc. We will examine their network centralities, such as hubs, bottlenecks and hierarchies, as we know that disruption of highly connected genes or their regulatory elements is more likely to be deleterious [13, 16]. In addition, we will employ probabilistic graphics models based framework to investigate influence of mutations and large structural alterations on various biological pathways. Furthermore, based on the pathway analysis, we will identify biological pathways, which are largely affected in IBC compared to non-IBC. This will help us to identify potential pathway level anomalies rather than recurrent single mutations to distinguish IBC from non-IBC samples. The Pusztai laboratory has also entered into a collaborative research agreement with IBM Genomics, which allows the team to beta test the Watson Genomics artificial intelligence software tool on the genomic data to interpret the multiplicity of anomalies in therapeutic and biological context. ***Genome-wide metrics:*** We will calculate and compare genome entropy, Mutant Allele Tumor Heterogeneity (MATH) and overall mutation load between the genomes of IBC and non-IBC to examine if a global metric of genome diversity defines the differences between IBC and non-IBC.

Aim 3. ***Validation of candidate molecular markers of IBC:*** We will validate candidate findings on an independent data set of 200 breast cancers including 60 IBC from the SWOG S0800 randomized clinical trial (http://clinicaltrials.gov/show/NCT00856492). DNA and RNA is currently being extracted from these samples in the Pusztai laboratory. Targeted DNA sequencing will be performed to estimate the prevalence of candidate genomic anomalies in the IBC and non-IBC cases. We will also have access to additional archived samples from the Yale Pathology Tissue Resource (YPTR) for further, focused validation of our findings.

Clinical Relevance

This study will generate the first whole genome sequence data on IBC, which on its own represents a significant contribution to the field by creating a data resource for the broader research community.

Discovery of a recurrent, IBC-specific genomic abnormality can lead to the development of an objective, molecular diagnostic criteria for IBC, similar to E-cadherin loss that defines invasive lobular carcinomas, and could standardize the diagnosis of this disease.

IBC is the most invasive form of breast cancer with local invasion that can progress within days and it also has a high propensity for distant metastatic spread. Identification of the genomic alterations associated with this aggressive phenotype could lead to new mechanistic insights into the biology of invasion and metastasis. Investigation of the biological function of the candidate genomic markers can be the subject of future grants and could spur new research directions in the laboratory.

Ultimately, identification of the genomic causes of IBC could lead to novel therapeutic strategies that may also have applications for other highly aggressive cancers.

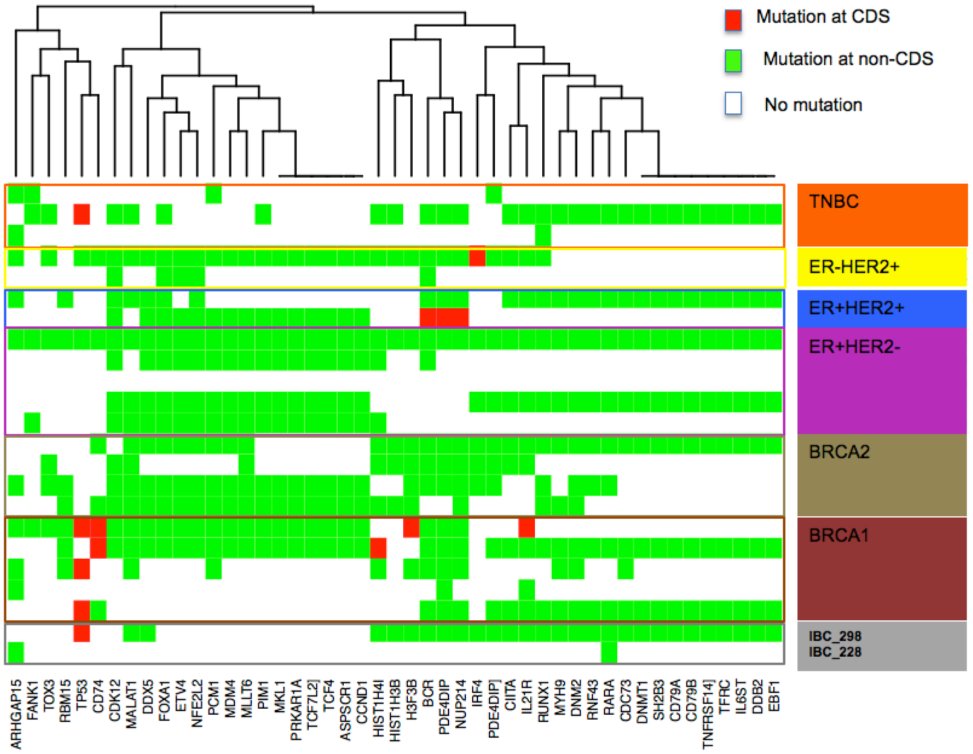

Figure 1. Genes with somatic variants prioritized by FunSeq2

Fig1 : All somatic variants (CDS and non-CDS regions) are prioritized by FunSeq2, and annotated with their related genes. Mutations common among all breast cancers influences genes associated with primary immunodeficiency or cancer pathways (listed above). However, Mutations that are on genes associated with P53 signaling pathway, are absent in IBC samples, but shared by all the other types of breast cancer samples.

SV Type	IBC_sample1 (germline)	IBC_sample1 (somatic)	IBC_sample2 (germline)	IBC_sample2 (somatic)
DEL	1362	9	1220	33
DUP	91	27	62	33
INV	1	0	0	0
ITX	36	26	39	13
CTX	37	7	32	16
Total	1527	69	1353	94

Table 1: Frequency of germline and somatic structural variations (SV) observed in 2 IBC samples. Each row represent different type of SV observed.

References

1. Masuda H, Baggerly KA, Wang Y, Iwamoto T, Brewer T, **Pusztai L**, Kai K, Kogawa T, Finetti P, Birnbaum D, Dirix L, Woodward WA, Reuben JM, Krishnamurthy S, Symmans WF, Van Laere SJ, Bertucci F, Hortobagyi GN and **Ueno NT**. Comparison of molecular subtype distribution in triple-negative inflammatory and non-inflammatory breast cancers. *Breast Cancer Research : BCR*. 2013;15:R112.
2. Iwamoto T, Bianchini G, Qi Y, Cristofanilli M, Lucci A, Woodward WA, Reuben JM, Matsuoka J, Gong Y, Krishnamurthy S, Valero V, Hortobagyi GN, Robertson F, Symmans WF, **Pusztai L, Ueno NT**. Different gene expression are associated with the different molecular subtypes of inflammatory breast cancer. Breast Cancer Res Treat 2011, 125(3):785-795.
3. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J,Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ,Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, Stephens PJ, McLaren S, Butler AP, Teague JW, Jönsson G, Garber JE, Silver D, Miron P,Fatima A, Boyault S, Langerød A, Tutt A, Martens JW, Aparicio SA, Borg Å, Salomon AV, Thomas G, Børresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ, Stratton MR and Breast Cancer Working Group of the International Cancer Genome Consortium. Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, *149*(5-10), 979–993.
4. Geraldine A. Van der Auwera, Mauricio O. Carneiro, Chris Hartl, Ryan Poplin, Guillermo Del Angel,Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Ki- ran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. From fastq data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics, 11(1110):11.10.1–11.10.33, Oct 2013.
5. Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., ... & Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*,*31*(3), 213-219.
6. Saunders, C. T., Wong, W. S., Swamy, S., Becq, J., Murray, L. J., & Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, *28*(14), 1811-1817.
7. Wang, J., Mullighan, C. G., Easton, J., Roberts, S., Heatley, S. L., Ma, J., ... & Zhang, J. (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nature methods*, *8*(8), 652-654.
8. Fu Y, Liu Z, Lou S, Bedford J, Mu X, Yip KY, Khurana E and **Gerstein M**. FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology*. 2014;15:480.
9. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M and **Gerstein M**. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biology*. 2012;13:R48.
10. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, Balasubramanian S, Beal K, Chakravarty D, Challis D, Chen Y, Clarke D, Clarke L, Cunningham F, Evani US, Flicek P, Fragoza R, Garrison E, Gibbs R, Gumus ZH, Herrero J, Kitabayashi N, Kong Y, Lage K, Liluashvili V, Lipkin SM, MacArthur DG, Marth G, Muzny D, Pers TH, Ritchie GR, Rosenfeld JA, Sisu C, Wei X, Wilson M, Xue Y, Yu F, Dermitzakis ET, Yu H, Rubin MA, Tyler-Smith C and **Gerstein M**. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*. 2013;342:1235587.
11. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT and McVean GA. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491:56-65.

12. Habegger, L., Balasubramanian, S., Chen, D. Z., Khurana, E., Sboner, A., Harmanci, A., ... & **Gerstein, M**. (2012). VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics*, *28*(17), 2267-2269.
13. Khurana, E., Fu, Y., Chen, J., & **Gerstein, M**. (2013). Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol*, *9*(3), e1002886.
14. TCGA consortium. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490:61-70.
15. Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., ... & Lichter, P. (2010). International network of cancer genome projects.*Nature*, *464*(7291), 993-998.
16. Kim, P. M., Korbel, J. O., & **Gerstein, M**. B. (2007). Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proceedings of the National Academy of Sciences*, *104*(51), 20274-20279.