Final Progress Report for Research Projects Funded by Health Research Grants

Instructions: Please complete all of the items as instructed. Do not delete instructions. Do not leave any items blank; responses must be provided for all items. If your response to an item is "None", please specify "None" as your response. "Not applicable" is not an acceptable response for any of the items. There is no limit to the length of your response to any question. Responses should be single-spaced, no smaller than 12-point type. The report **must be completed using MS Word**. Submitted reports must be Word documents; they should not be converted to pdf format. Questions? Contact Health Research Program staff at 717-783-2548.

1. **Grantee Institution:** NSABP Foundation, Inc.

2. **Reporting Period (start and end date of grant award period):** 1/1/2009 to 12/31/2011

3. **Grant Contact Person (First Name, M.I., Last Name, Degrees):** Donna Szczepankowski

4. **Grant Contact Person's Telephone Number:** 412-330-4610

5. **Grant ME Number or SAP Number:** 4100047643

6. **Project Number and Title of Research Project:** Development of Prognostic Index for Colon Cancer Patients Using Gene Expression Profiling

7. **Start and End Date of Research Project:** 1/1/2009 to 12/31/2011

8. **Name of Principal Investigator for the Research Project:** Soonmyung Paik, MD

9. **Research Project Expenses**.

   9(A) Please provide the amount of health research grant funds spent on this project for the entire duration of the grant, including any interest earned that was spent:

   $ 1,281,815.58

   9(B) Provide the names of **all** persons who worked on this research project and were supported with health research funds. Include position titles (Principal Investigator, Graduate Assistant, Post-doctoral Fellow, etc.), percent of effort on project and total health research funds expended for the position. For multiple year projects, if percent of effort varied from year to year, report in the % of Effort column the effort by year 1, 2, 3, etc. of the project (x% Yr 1; z% Yr 2-3).

Last Name	Position Title	% of Effort on Project	Cost
Paik	Principal Investigator; Director of the Division of Pathology	3% Yr 1; 8% Yr 2-3	$26,192
C. Kim	Manager, Molecular Pathology Laboratory	35% Yr 3	$27,961
S. Kim	Pathology Research Fellow	30% Yr 3	$15,564
Pogue-Geile	Assistant Director of Molecular Profiling	5% Yr 2; 30% Yr 3	$21,461
Remillard	Research Assistant	5% Yr 1; 15% Yr 2	$12,680
Blackmon	Research Assistant	5% Yr 1; 40% Yr 2-3	$32,895
Lipchik	Research Assistant	5% Yr 2; 40% Yr 3	$16,994
Yamaguchi-Tanaka	Biostatistician/Computational Biologist/Bioinformatics Specialist	100% Yr 3	$76,692
Gavin	Molecular Technical Supervisor	4% Yr 2; 40% Yr 3	$20,022
Oeler	Senior Research Histologist	8% Yr 2-3	$ 7,611
Reilly	Research Assistant	25% Yr 2-3	$ 2,218
Bosshart	Office Assistant	2% Yr 3	$ 971

9(C) Provide the names of **all** persons who worked on this research project, but who *were not* supported with health research funds. Include position titles (Research Assistant, Administrative Assistant, etc.) and percent of effort on project. For multiple year projects, if percent of effort varied from year to year, report in the % of Effort column the effort by year 1, 2, 3, etc. of the project (x% Yr 1; z% Yr 2-3).

Last Name	Position Title	% of Effort on Project
Szczepankowski	Director of Finance & Sponsored Projects	2%
Cortina	Grant & Special Projects Administrator	5%
Costantino	Biostatistician	2%

9(D) Provide a list of **all** scientific equipment purchased as part of this research grant, a short description of the value (benefit) derived by the institution from this equipment, and the cost of the equipment.

Type of Scientific Equipment	Value Derived	Cost
NCounter Analysis System	Provided a simple methodology and platform for generation of gene expression data from for RNAs isolated from FFPET	$ 211,500

10. Co-funding of Research Project during Health Research Grant Award Period. Did this research project receive funding from any other source during the project period when it was supported by the health research grant?

Yes_____ No_____X_____

If yes, please indicate the source and amount of other funds:

11. Leveraging of Additional Funds

11(A) As a result of the health research funds provided for this research project, were you able to apply for and/or obtain funding from other sources to continue or expand the research?

Yes_____ No__X_____

If yes, please list the applications submitted (column A), the funding agency (National Institutes of Health—NIH, or other source in column B), the month and year when the application was submitted (column C), and the amount of funds requested (column D). If you have received a notice that the grant will be funded, please indicate the amount of funds to be awarded (column E). If the grant was not funded, insert "not funded" in column E.

Do not include funding from your own institution or from CURE (tobacco settlement funds). Do not include grants submitted prior to the start date of the grant as shown in Question 2. If you list grants submitted within 1-6 months of the start date of this grant, add a statement below the table indicating how the data/results from this project were used to secure that grant.

A. Title of research project on grant application	B. Funding agency (check those that apply)	C. Month and Year Submitted	D. Amount of funds requested:	E. Amount of funds to be awarded:
None	☐NIH ☐ Other federal (specify:_____) ☐ Nonfederal source (specify:_)		$	$

11(B) Are you planning to apply for additional funding in the future to continue or expand the research?

Yes____X_____ No_____

If yes, please describe your plans:

We will continue to identify the best model for prognostication and prediction for oxaliplatin benefit. Currently we are doing exploratory analysis with the C-07 discovery cohort (N=866) utilizing the nCounter or NanoString data, which consists of gene expression profiles from 282 prognostic or oxaliplatin-predictive genes. Detailed protocols for model building and their evaluation using the C-07 nCounter assays have been developed and are described below. We plan to carry out these analyses and to carry out validation of models with the remaining 918 cases which have been set aside as the validation cohort. The gene expression profiles for these cases have already been completed.

12. Future of Research Project. What are the future plans for this research project?

The future plans for this research project are to leverage the discoveries for additional funding from NIH or other funding sources such as AACR or ACS. Once we have validated a prognostic model within the C-07 data, we will seek additional funds to validate the model using NSABP clinical trial C-06. NSABP clinical trial C-06 compared oral uracil/Ftorafur (UFT) plus leucovorin (LV) to 5-fluorouracil (5-FU) Plus LV in the treatment of patients with Stages II or III colon cancer.

Prediction of treatment is particularly difficult and so our goal of defining a model that would predict benefit from oxaliplatin may not be possible; however, if a prediction model is validated with the C-07 validation cohort, we will seek collaborations with other investigators to validate our predictive model. Another approach we may use to identify a signature that could be used to help to identify patients who are most appropriate for receiving benefit from use of oxaliplatin is to develop a prognostic model in patients who are considered to be high risk and who are often treated with oxaliplatin. Dividing into subsets the prognoses of patients who in the clinical setting will receive oxaliplatin treatment would be clinically useful. Even though the signature may not demonstrate oxaliplatin interaction, it may identify low-risk patients whose prognoses are so good that they are unlikely to receive any benefit from oxaliplatin.

13. New Investigator Training and Development. Did students participate in project supported internships or graduate or post-graduate training for at least one semester or one summer?

Yes___X_____ No_____

If yes, how many students? Please specify in the tables below:

	Undergraduate	Masters	Pre-doc	Post-doc
Male				
Female				1
Unknown				
Total				**1**

	Undergraduate	Masters	Pre-doc	Post-doc
Hispanic				
Non-Hispanic				1
Unknown				
Total				**1**

	Undergraduate	Masters	Pre-doc	Post-doc
White				
Black				
Asian				1
Other				
Unknown				
Total				**1**

14. Recruitment of Out-of–State Researchers. Did you bring researchers into Pennsylvania to carry out this research project?

Yes____X_____ No_____

If yes, please list the name and degree of each researcher and his/her previous affiliation:

Matthew Remillard, B.S., recent graduate of Carnegie Mellon University, resident of Massachusetts
Seong-Rim Kim, M.D., Medplan Pathology Laboratory Center/ Samsung Medical Center, Seoul, S. Korea
Noriko Yamaguchi-Tanaka, Ph.D., Dana Farber Cancer Institute and Harvard School of Public Health, Boston, MA

15. Impact on Research Capacity and Quality. Did the health research project enhance the quality and/or capacity of research at your institution?

Yes__X_____ No_____

If yes, describe how improvements in infrastructure, the addition of new investigators, and other resources have led to more and better research.

Research was enhanced by the addition of new investigators. The following investigators brought unique skills and or education for the fulfillment of this project. Seong-Rim Kim, trained as a pathologist in Seoul, South Korea. For this project, she defined tumor tissue so that we were able to isolate RNA from samples that were composed of mostly tumor cells. Matthew Remillard, a graduate from Carnegie Mellon University, and now a graduate student in the Department of Molecular and Cellular Biology at Princeton University, brought not only good molecular biological skills to the project, but he also had excellent IT skills. He was instrumental in automating our RNA-isolation process. Patrick Gavin, who was previously a Production Associate responsible for high throughput SNP genotyping with MassArray at Sequenom was instrumental in automating the DASL® and nCounter gene-expression processes, and in meticulously processing the gene expression data for these platforms. Noriko Yamaguchi, recruited from the Dana Farber Cancer Institute and the Harvard School of Public Health, was critical to the development of a protocol for the development of models for prognosis and prediction. Her training and expertise in biostatistics and bioinformatics makes her well suited to performing the analysis of our large, expression data set from NSABP clinical trial C-07.

We have improved our method of RNA extraction through the funding provided by this project. The throughput for RNA extraction has vastly improved by development of a method that allows for the purification of RNA from 96 sample lysates in one hour. Previously, the throughput for the isolation of RNA from sample lysates was 16 samples per hour.

Purchase of the nCounter system vastly improved our research capabilities for gene expression analysis. The nCounter assays are an ideal platform for the analysis of RNAs isolated from routinely collected, formalin-fixed, paraffin-embedded tissue (FFPET) samples. RNAs from FFPET are degraded and chemically modified. The nCounter assays involve the hybridization of the RNA to a set of multiplexed custom oligonucleotides, does not involve any enzymatic reactions, and requires RNA sizes of only 100 nucleotides or less. This makes the nCounter assays ideally suited as assays for RNAs isolated from FFPET, and, also, the simplicity of the assays makes them an ideal clinical assay.

16. Collaboration, business and community involvement.

16(A) Did the health research funds lead to collaboration with research partners outside of your institution (e.g., entire university, entire hospital system)?

Yes_____ No_____X_____

If yes, please describe the collaborations:

16(B) Did the research project result in commercial development of any research products?

Yes_____ No____X_____

If yes, please describe commercial development activities that resulted from the research project:

16(C) Did the research lead to new involvement with the community?

Yes_____ No____X_____

If yes, please describe involvement with community groups that resulted from the research project:

17. Progress in Achieving Research Goals, Objectives and Aims.

List the project goals, objectives and specific aims (as contained in the grant application's strategic plan). Summarize the progress made in achieving these goals, objectives and aims <u>for the entire grant award period.</u> Indicate whether or not each goal/objective/aim was achieved; if something was not achieved, note the reasons why. Describe the methods used. If changes were made to the research goals/objectives/aims, methods, design or timeline since the original grant application was submitted, please describe the changes. <u>Provide detailed results of the project.</u> Include evidence of the data that was generated and analyzed, and provide tables, graphs, and figures of the data.

Note: This response should be a DETAILED report of the methods and findings. It is not sufficient to state that the work was completed. Insufficient information may result in an unfavorable performance review, which may jeopardize future funding. **There is no limit to the length of your response. Responses must be single-spaced below, no smaller than 12-point type. If you cut and paste text from a publication, be sure symbols print properly, e.g., the Greek symbol for alpha (α) and beta (ß) should not print as boxes (□). DO NOT DELETE THESE INSTRUCTIONS.**

Summary of Goals and Specific Aims
The broad research objective of this project was to improve the clinical care of patients with colon cancer diagnoses. The specific aim was to develop a prognostic test using gene expression profiling to identify patients who may not require more than 5-fluorouracil plus leucovorin (FULV) adjuvant chemotherapy. The current standard of care for Stage II or III colon cancer is adjuvant chemotherapy with an oxaliplatin-containing regimen (FULV plus oxaliplatin [FLOX or FOLFOX, based on schedule]). This standard was established by the NSABP trial C-07, which showed the superiority of FLOX over FULV. However, oxaliplatin regimens are very toxic with numerous side effects, including neurotoxicity. Data from NSABP C-07 and preceding trials conducted by our group and others make it clear that many patients may not require oxaliplatin. Yet, there are no reliable prognostic markers in clinical use to identify patients whose prognoses are good enough after treatment with only FULV such that oxaliplatin is not required. Developing such a prognostic test will relieve suffering from unnecessary toxic therapy and lead the way to personalized care of colon cancer patients.

The Specific Aims were:
1) Discover candidate prognostic genes for FULV-treated colon cancer patients by gene expression profiling of tumor blocks from NSABP trial C-07 using a whole genome DASL arrays from Illumina®
2) Design the nCounter assay for candidate genes and assay the NSABP C-07 cohort again to build a prognostic algorithm; and
3) Prospectively validate the prognostic algorithm in independent cohorts of FULV-treated patients from NSABP trials C-05 and C-06 using the nCounter assay.

1) *Discover candidate prognostic genes for FULV-treated colon cancer patients by gene expression profiling of tumor blocks from NSABP trial C-07 using a whole genome DASL arrays*

Expression profiling with DASL. Whole genome DASL arrays from Illumina were used to profile the gene expression of approximately ½ (N=956) of the total number tumor blocks (N=2222) collected from patients enrolled in NSABP clinical trial C-07. However, we discovered that it was necessary to refine our discovery cohort slightly and as a result a new discovery cohort was defined which consisted of 866 cases.

Prognostic models were built with C-07 DASL whole genome expression data.
We initially attempted to build a model using the oxaliplatin arm of the trial because we wanted to validate the model with cases from the control arm of NSABP C-08 (modified FOLFOX6

regimen). In this way, both the discovery and validation data sets would consist of patients that were both treated with nearly identical oxaliplatin regimens. However, when we attempted to build models in the oxaliplatin arm of the C-07, the number of prognostic genes in the oxaliplatin arm of C-07 was very small and the models to distinguish high- and low-risk patients were less satisfactory than models in which only the clinical covariates were used in the model. Conversely, building models with the gene expression data with only the control arm (FULV) of C-07 resulted in good prognostic models, meaning that high- and low-risk groups achieved some additional separation when added to clinical covariates in the model.

A prognostic model was built using both the FULV and FLOX arms of C-07, using genes with a p value less than 0.00004 which corresponded to a false discovery rate (FDR)<0.1. Additionally, genes were filtered to ensure that 20% of the expression data for a particular gene had at least a 1.5-fold change in either direction from the gene's median value. Supervised Principal Component analysis (SuperPC) as a part of BRB (Biometric Research Branch)-Array Tools was used to build a model using 3 principle components and 10-fold cross-validation. The model that included clinical covariates for treatment, T-stage, and nodal status was built with 20 genes (Fig.1). The gene expression model was able to correctly classify patients into high- and low-risk groups and did add some additional separation to high- and low-risk patients identified with only clinical variables included in the model.

The prognostic model, which was built using samples from both arms of the C-07 trial and whole-genome expression DASL data, was tested in cases from C-08. Figure 2 demonstrates that the 20-gene prognostic model built with C-07 DASL data was validated in the C-08 DASL data (Fig 2). However, a comparison of the models built with only clinical covariates to models built with clinical covariates plus genes demonstrates that the gene expression did not improve the prognostic model (Fig 3).

While the building of gene expression models with the DASL data was an important exercise for understanding how this data might be able to distinguish high- and low-risk patients, the ultimate goal of this work is to be able to devise a clinical test that could be used to determine which patients can benefit from oxaliplatin. Because whole genome expression profiling is too complex and too expensive to be used as a clinical test, we decided to focus on the selection of prognostic and predictive genes for profiling on a more clinically amenable platform.

2) *Design the nCounter assay for candidate genes and assay the NSABP C-07 cohort again to build a prognostic algorithm*

Selection of prognostic and predictive genes. Prognostic and predictive genes using the DASL data were selected for the purpose of shifting the expression analysis to the nCounter platform. The nCounter system by NanoString is an ideal system for clinical evaluation of RNA because it involves the hybridization of the RNA to a set of multiplexed custom oligonucleotides, does not involve any enzymatic reactions, and requires small amounts and sizes of RNA. This makes the nCounter system ideally suited not only as platform for interrogation of RNAs isolated from FFPET but also as clinical test due to the simplicity of the assays. Therefore, we set out to select prognostic and predictive genes for inclusion into the nCounter code set to be used for prognosis

and for prediction of oxaliplatin benefit. We refer to this code set as the nCounter Colon code set3.

Prognostic genes dependent and independent of clinical variables were identified by calculating the non-adjusted (crude) and adjusted HR and p-values for each of the 24,525 probes present on the Illumina whole-genome DASL arrays. Predictive genes for oxaliplatin benefit were identified by calculation of interaction Cox proportional hazard model including treatment-by-gene expression interaction terms for each gene.

Predictive genes for benefit for oxaliplatin benefit were identified by calculation of interaction via a Cox proportional hazard model including treatment-by-gene expression interaction terms for each gene. Sixty-nine genes were identified with an interaction p-value below 0.01 which were used as a cut off to identify predictive genes. Predictive genes were also identified by 10-fold cross validation and 39 genes were identified with an average p value below 0.01. A Kaplan-Meier plot in which patients are stratified by the expression levels (using a median cut) of one of these predictive genes and by treatment is shown in Figure 4. Clearly, patients with a low level of expression of this gene appear to receive benefit from oxaliplatin while patients with a high level of expression do not. Several such genes were identified that showed a similar pattern in Kaplan-Meier plots.

Exploratory analysis of the C-07 DASL data also identified prognostic genes by implementing survival pathway analysis within BRB-Array Tools. Analysis of the C-07 DASL data identified molecular pathways that were associated with survival. We observed as a result of this exploratory analysis that there were distinct differences in the pathways identified in the control arm (FULV) compared to the test arm (FLOX). This observation was consistent with our observation that there was very little overlap in the prognostic genes identified in the control and oxaliplatin arms.

Additional criteria were generated to evaluate the candidate genes because we had previously observed that gene expression profiling results obtained with whole-genome arrays were often discordant from those obtained with other smaller platforms such as the nCounter system. We observed that genes with a small dynamic range or with very low expression values did not readily validate in the second platform. Therefore, we performed additional data analysis beyond the examination of the prognostic or predictive values from the DASL data in order to select genes for the nCounter Colon code set3. We calculated the interquartile ranges for the candidate prognostic genes.

Selection of prognostic genes from the C-07 DASL data were based on the following criteria:

1. prognostic in the control arm or in the combined arms of the trial at $p<0.001$ (most of these genes had an interquartile range [IQR] >1.2), or
2. prognostic genes identified by analysis of survival pathways via BRB-Array Tools in the oxaliplatin arm, the control arm, or the combined arms and with a prognostic $p<0.01$.

The criteria for the selection of genes for predicting oxaliplatin benefit were based on interaction tests using 10 fold-cross validation (CV). The predictive genes had the following statistics:

1. 10xCV $p<0.01$, IQR >0.8. or

2. p-values $p<0.05$, IQR >1.2, median>8, and, in addition, Kaplan-Meier plots for these genes suggested that these genes could be useful for determining oxaliplatin benefit or the function of this gene suggested it could influence oxaliplatin response.

Other prognostic and predictive genes were included in nCounter Colon code set3. They included genes that had previously been identified for a different project in which the nCounter code set1 was used to profile cases from NSABP clinical trial C-08. In this trial, the control arm was nearly identical to the test arm of C-07. Therefore, we selected prognostic genes with $p<0.001$from the oxaliplatin arm of C-08. We also included a few genes from the literature which had been reported to be prognostic and or predictive for oxaliplatin. In addition, gene expression data available in the Gene Expression Omnibus (GEO) database were analyzed and genes in which multiple probes were found to be significant at $p< 0.0001$ were included in the code set.

When the above criteria were applied, 359 probes or genes were identified. The genes from the DASL data were identified with the Illumina ID, because we wanted to use the same region of the gene as had been interrogated with the probe from the DASL data for the design of the nCounter assays in order to enhance the possibility of replicating the DASL data results. Genes identified by BRB-Array Tools or from the literature were identified by the official gene name, rather than the probe name. When the Illumina probes were annotated and updated for their current assignment to a gene, it was found that the actual number of different genes described by the 359 probes and genes was reduced to 282. The assignment of a specific sequence to a gene is a dynamic process. As a result, the number of genes and the sequences assigned to genes are altered every time improvements are made in the build of the human genome. The most recent build of the human genome is 19 (GRCh37). Probes on the Illumina DASL array were designed based on a build of the genome from approximately 2006 when the array was built, but in 2011, some of the probes had been reassigned to a different gene.

In addition to the 282 prognostic and predictive genes, 14 housekeeping genes for normalization of the arrays were included by us, and 14 positive and negative controls are included by NanoString as quality controls for the hybridization.

Assay the NSABP C-07 cohort with the nCounter colon code set3. The nCounter system from NanoString is a digital detection system capable of highly multiplexed direct profiling of individual molecules without amplification. The nCounter assays include a reporter probe and capture probe for each gene. The reporter probe consists of a color coded bar code attached to a specific oligonucleotide homologous to the gene of interest. The capture probe contains a molecule that binds the NanoString cassette and an oligonucleotide homologous to the same gene as the reporter probe but located approximately 100 nucleotides distant from the reporter probe sequence. The system involves the hybridization of many multiplexed reporter and capture probe pairs to the RNA, the transfer of the samples to a NanoString prep station where excess probes are removed, and probe/target complexes are bound, immobilized and aligned on a NanoString cassette. The cassette is then loaded into the digital analyzer and the bar codes are counted for each target molecule. This whole process requires the pipetting of 3 reagents/sample, and hands-on time of approximately 1 hour over a day-and-a-half to process 72 samples with as many as 800 genes.

We have profiled both the discovery (N=860) and validation cohorts (N=915) with the 310 genes included in the nCounter Colon code set3. We chose to profile both the discovery and validation code sets at one time to avoid complications due to batch effects. We have previously observed that profiling samples with 2 code sets composed of identical genes but that were manufactured at 2 different times resulted in batch effects that complicated analysis.

Build a prognostic algorithm. The C-07 nCounter dataset represents a very large number of samples and the prognostic and predictive genes have been carefully chosen. Analysis of the discovery cohort will involve a comprehensive exploratory analysis in order to identify the best possible model that has the best chance of being validated in the validation cohort. Therefore, considerable care and time has been invested in developing a detailed protocol describing the statistical methods that will be used to identify prognostic and predictive models. Therefore, we have included the details of our protocol for the analysis of the nCounter discovery cohort of C-07.

Protocol for the Statistical Methods for the Building of Prognostic and Predictive Models
Because we observed in C-07 DASL data that prognostic genes were distinctly different between the control arm (FULV) and the experimental (FLOX) arm, we proposed to look at each arm separately and together.

Prognostic models
Model 0 will be built as a prognostic model using only the information of treatment and TNM staging. Model 1-1 will be a prognostic model based on univariate selection. Each gene will be evaluated by applying univariate Cox regression models, and the top *k* genes with p-values less than 0.05 will be selected. Furthermore, a multivariate Cox regression will be used to select by applying forward selection with clinical covariates and *k* genes. Model 1-2 will be built using Supervised Principal Component (SuperPC) analysis using the information of genes and other clinical covariates (treatment, sex, age, TNM staging). Genes identified as meeting the FDR rate of 0.1 or p-value from the Cox model of less than 0.05 in the primary analysis will be used in this model. The criteria to screen genes are dependent upon the number of genes that are selected in the next stage. We will apply SuperPC as part of BRB-Array Tools to C-07 samples using nCounter assay data for the genes meeting entry criteria. The first 3 principle components will be used in the models. Model 1-3 which will be built with Cox univariate shrinkage (CUS). The R package uniCox will be used to build the model using CUS. Clinical covariates will be included in the model. Model 1-4 will be built with Random Forests™ (RF). R package randomSurvivalForest will be used to build the model using RF. Clinical covariates will be included in the model.

Model 2 will be built using biological information. Previous analysis of C-07 whole-genome expression utilizing DASL data identified genes associated with immune reaction and cytotoxic T cells as prognostic genes almost exclusively in the control (FULV) arm (along with genes associated with apoptosis and cell division). In the oxaliplatin arm, prognostic genes were mostly related to cell division, apoptosis, membrane transport, and metabolism. Thus, we anticipate that prognostic genes in the C-07 control arm will be related to immune response, including a cytotoxic T-cell response, apoptosis, and cell division and will be preferentially

selected for model building. For the oxaliplatin arm, we anticipate that prognostic genes will play a role in cell division, absorption, oxaliplatin metabolism, and apoptosis and will be preferentially selected for model building.

1. Patient subgroups will be defined using *k*-means cluster analysis.
2. Stepwise discriminant analysis will be used to select a subset of the quantitative variables for use in discriminating among the subgroups.
3. The canonical discriminant function will be estimated to predict each patient's subgroup. The number of the canonical discriminant function will be defined based on the amount of a given value that explains >70% of the data or eigen values that are larger than 1. We will build the prognostic model using the Cox regression model with canonical correlation score calculated with the genes selected in step 2.
4. For the validation set, we will apply both the canonical discriminant function and parameters in the Cox regression model estimated from the discovery set to calculate the deviance of the Cox model.

Method assessment of prognostic models.
We will apply 10-fold cross-validation to build each model. To evaluate the models, averaged deviance will be calculated for each model and the model with the smallest averaged deviance will be chosen for the validation. If two methods have comparably small average squared deviance, we can use other factors to select one.

Plans for building oxaliplatin predictive models.
The goal is to provide a model that will improve on the current clinical guidelines for oxaliplatin treatment recommendations. In order to understand whether the prediction models that are built with gene expression information can improve upon how patients are selected for oxaliplatin treatment, we will build a model using only clinical co-variates which will include only Stage III patients and Stage II patients if just one of the following: lymphovascular invasion or perforation or T4 lesion or grade 3-4 lesions or bowel obstruction at presentation.

Model 0 refers to the model based strictly on the above clinical variables. Based on this definition of subgrouping, we will estimate the parameter from the Cox linear regression model with an interaction term of the above indicator variable with treatment in the training set and will calculate averaged deviance and the interaction effect from the test set.

Several different prediction models for oxaliplatin benefit will also be built and evaluated using the nCounter data. We will apply 10-fold cross-validation to build each model. All statistics calculated to build predictive model will be adjusted for clinical covariates (sex, age, TNM stage).

Model 1-1 will be built based on univariate p-value assuming linear interaction effect. Genes will be discovered within the training set with p-value below 0.05 for interaction of genes and treatment in a linear model, or the top 21 (target 282 genes *0.05) genes with smaller p-values will be regarded as candidate genes if there are more than 21 genes with p-values <0.05. Then we will pick out the top *k*-ranked selected genes using backward elimination.

Model 1-2 will be built based on univariate p-value calculated by STEPP (Subpopulation Treatment Effect Patient Plot) analysis and categorical analysis assuming the existence of non-linear interaction effects. Discovery genes within the training set with p-value below 0.05 for interaction of genes and treatment calculated with STEPP method and the Cox regression model with categorized expression values, or top 21 genes with smaller p-values will be regarded as candidate genes if there are more than 21 genes with p-values <0.05. As well as model 1-1, the *k* top ranked selected genes will all be included in a multivariate COX regression model using backward elimination, but the interaction effect for each gene will be assessed with optimal transformation based on STEPP analysis for each candidate gene.

Additional models for both prognosis and prediction are currently being evaluated and developed. Methods for model assessment have been detailed in our protocol. All the relevant investigators will agree on the model that will be used to validate the model by analyzing the data in the validation cohort which consists of cases from C-07 which were not used in the development of the model. Because only one model can be tested in the validation cohort, it is essential that careful analyses of the different models are conducted so that we make the best choice possible for validation.

Detailed Methods

Isolation of RNA. RNAs were isolated for the discovery cohort using four 5-micron slides, utilizing a semi-automated RNA isolation method that required 2 days and approximately 8 hours of hands-on time. Total RNA was extracted from the first discovery subset of 1,080 samples from the NSABP C-07 clinical trial. On the first day, four 5-µm unstained slides were cut from the formalin-fixed, paraffin-embedded (FFPE) tumor blocks. Tumor areas were identified on an H&E slide by a pathologist and corresponding sections were isolated from the unstained slides by macrodissection. A protease digestion was then performed overnight. Extractions were performed using the magnetic particle processor KingFisher®ml (Thermo Fisher Scientific Inc, Waltham, MA) in combination with the MagMAX™ Total RNA Isolation Kit (Applied Biosystems, Foster City, CA). The microspherical magnetic beads provided by the kit and used in this procedure have a large available binding surface and can be fully dispersed in solution, allowing thorough RNA binding, washing and elution. The Quant-iT™ RiboGreen® RNA Assay Kit (Invitrogen, Carlsbad, CA) and the Infinite® F200 fluorometer (Tecan, Mannedorf, Switzerland) was used to quantitate the RNA. The total yield obtained ranges from 500 ng to 19 µg (average: 4.5 µg). This allows for the purification of 16 RNAs from lysates in approximately an hour.

We improved our method of RNA extraction for the validation cohort this year by increasing the throughput by switching to instruments that allow 96 samples to be processed in one hour. Approximately, 1000 additional RNAs from the C-07 validation cohort have been prepared in this way. Specifically, RNAs were isolated from lysates prepared from 3 slides at a width of 5 µm. Lysates were deparaffinized and underwent protein digestion and then were loaded into a 96-well plate with the TeCan Evo robot and processed with the King FisherFlex 96 instrument (ThermoFisher, Burlington, ON) and the E.Z.N.A.® FFPE RNA Isolation Kit from Omega Bio-Tek kit (Norcross, GA). RNAs were quantified with fluorescence, using the Quant-iT ™ RiboGreen® Assay Kit (Invitrogen, Carlsbad, CA) and the InfiniteF200 fluorometer (Tecan, Mannedorf, Switzerland).

RNAs (100-200 ng) from 971 samples were profiled for gene expression utilizing the whole-genome DASL (cDNA-mediated Annealing, Selection, extension, and ligation) Arrays from Illumina (San Diego, CA). DASL was specifically designed to interrogate partially degraded RNAs and was, therefore, ideal for samples isolated from FFPET. The details of this procedure are provided by the manufacturer. Briefly, DASL involves cDNA synthesis, hybridization of a complex set of oligonucleotides that allow for the synthesis and amplification of more than 24,000 annotated genes derived from RefSeq (Build 36.2). Expression is interrogated by direct hybridization to Illumina HumanRef-8 bead arrays. After hybridization, BeadChips were scanned using Illumina Bead Array Reader, images were digitized and data were collected and viewed within Illumina's GenomeStudio® Data Analysis Software package and exported to third party software for further analysis.

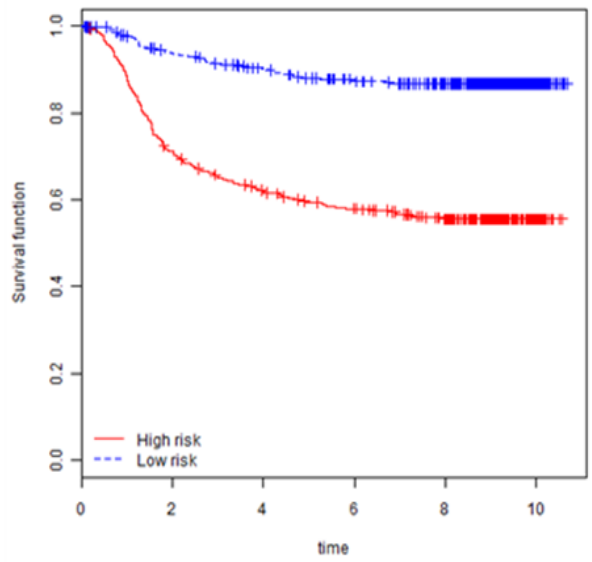
Immunohistochemistry Staining for Mismatch Repair Proteins Tissue microarrays were created and immunostaining was performed for microsatellite instability markers, including MSH2, MLH1, and PMS2. Antibodies for immunohistochemistry included anti-MSH2 (clone FE-11; 1:200; Calbiochem, Darmstadt, Germany) and anti-MLH1 (clone G168-15; 1:30; BD Biosciences, San Jose, CA USA). Sections were incubated with normal goat serum for 30 minutes and then with specific mouse monoclonal antibodies overnight at 4°C and stained with DAKO EnVision ™ + system Kit (DAKO) according to the manufacturer's instructions. Mismatch repair status was determined by the staining results for MSH2 and MLH. Tumors that showed no staining with anti-MSH2 or anti-MLH1 or both in the tumor cells were considered to be MMR-defective or unstable. Only tumors that showed staining for both were considered to be MMR-proficient or stable. Positive staining in non-tumor cells was used as a control for assay success.

Summary of Work Completed

In summary, we have successfully carried out whole-genome expression analysis using DASL arrays using a large cohort of tumors from patients enrolled in NSABP clinical trial C-07. We have demonstrated that prognostic genes and models can be developed with the DASL data but did not pursue the models because we thought that it was more important to pursue models with a gene expression platform that was amenable to clinical evaluation. Therefore, we identified a large number of prognostic and predictive genes and designed an nCounter assay—because the nCounter system is simple and cost-effective and highly concordant with RT-PCR, which is considered the gold standard for expression analysis. We have completed gene expression analysis of 282 candidate prognostic and oxaliplatin-predictive genes on 1775 cases from NSABP C-07. We have developed a comprehensive protocol for the development and evaluation of prognostic and predictive models that may improve upon current methods that are used in clinical practice to determine prognosis and to determine whether patients should be treated with oxaliplatin. Currently, we are in the process of building and evaluating these models and will pursue additional funding for signature evaluation. In conclusion, we have made significant progress toward our goal identifying genes that can improve upon the prognosis and treatment of colon cancer.

Fig 1. 20 Gene Model

Model Fitted with Genes and Covariates



Model Fitted with Covariates Only

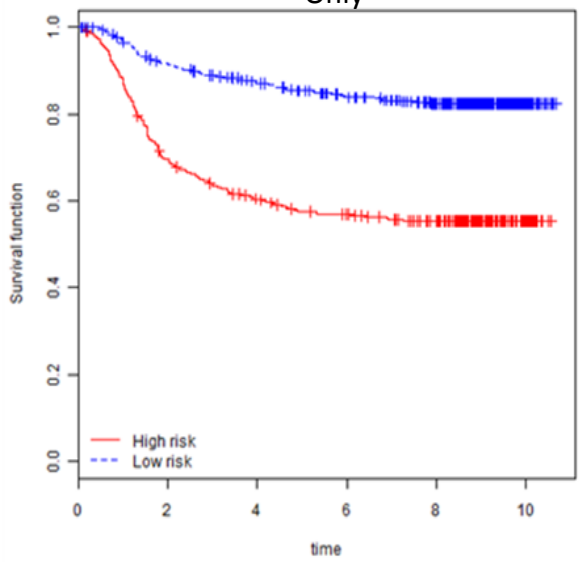


Fig. 2. 20 Gene Model is Validated in C08

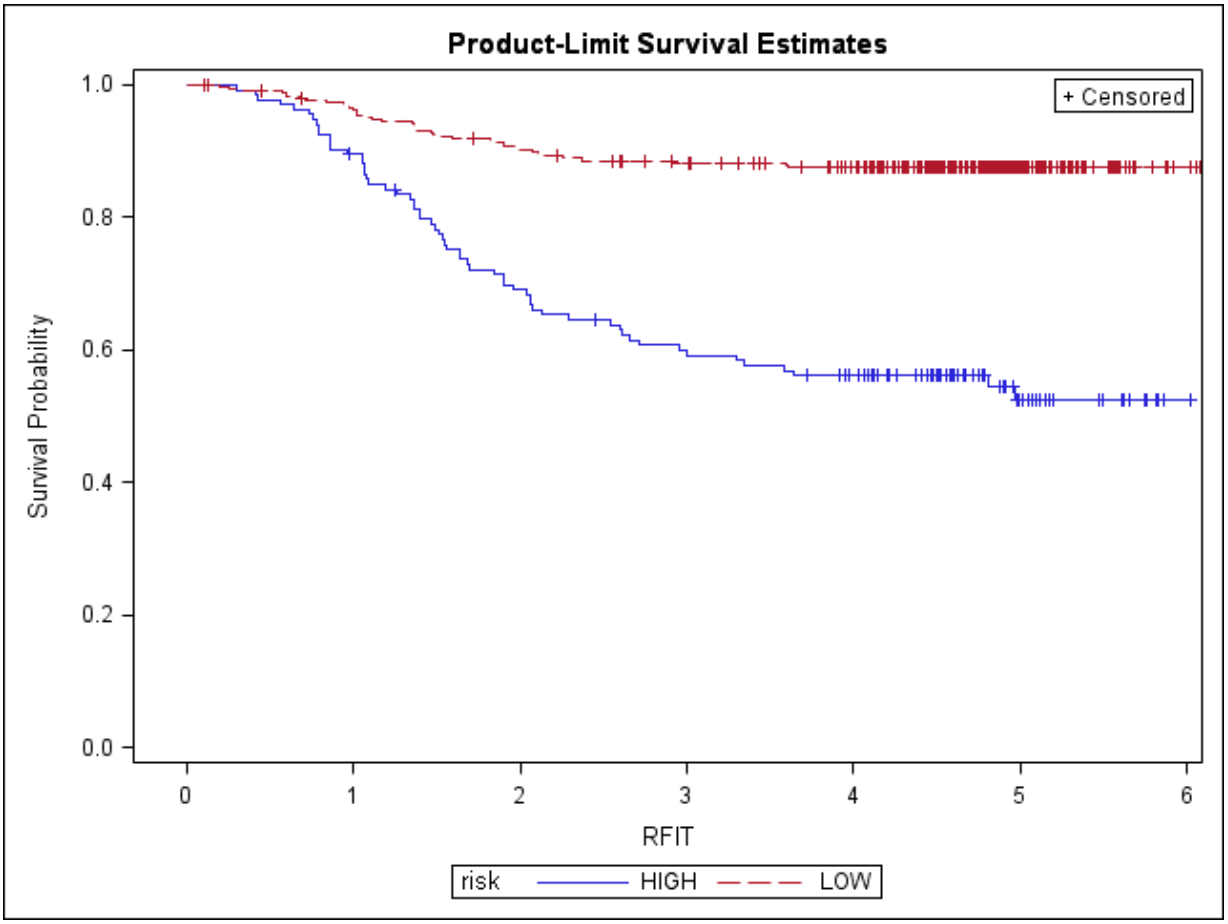
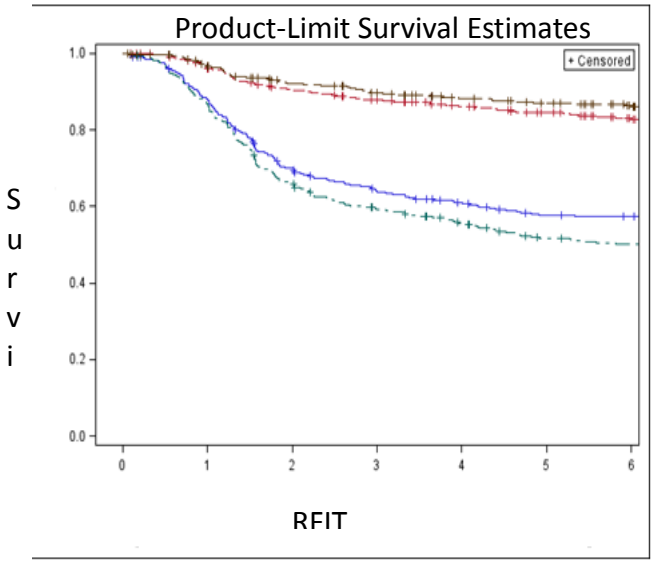
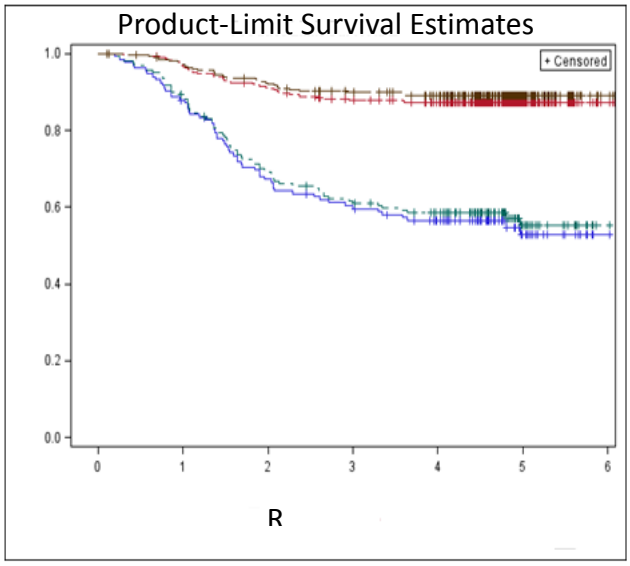


Fig 3. Kaplan-Meier Comparisons of Clinical and Gene Models

C-07



C-08



S
u
r
v
i

C-index for COV only model=0.62
C-index for COV for COV + gene model=0.65

C-index for COV only model=0.63
C-index for COV for COV + gene model=0.65

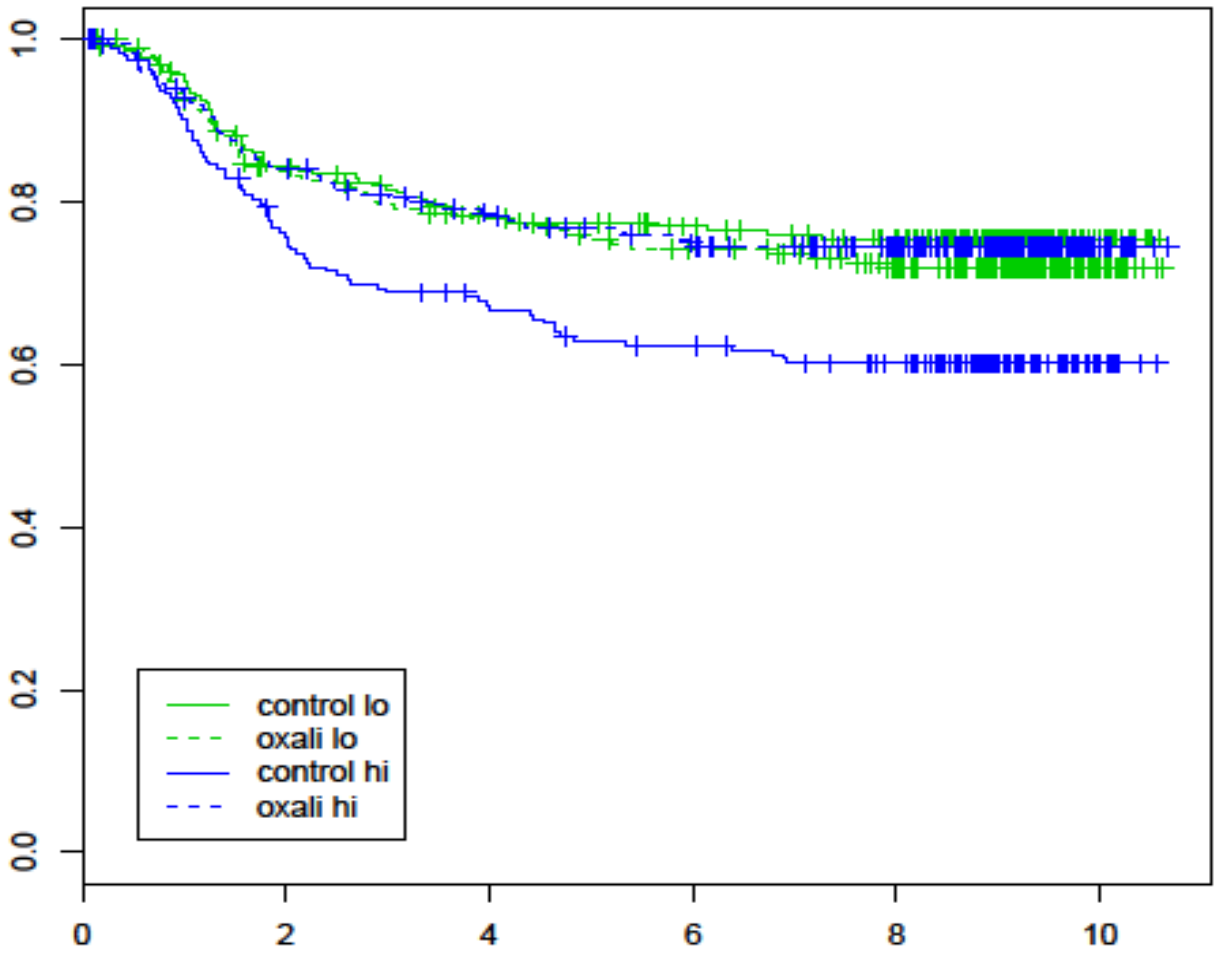
Model= COV + genes, low risk

Model= COV, low risk

Model= COV, risk high

Model COV+ genes, high risk

Fig 4. Gene Expression from a single gene demonstrates predictive value from oxaliplatin treatment



18. Extent of Clinical Activities Initiated and Completed. Items 18(A) and 18(B) should be completed for all research projects. If the project was restricted to secondary analysis of clinical data or data analysis of clinical research, then responses to 18(A) and 18(B) should be "No."

18(A) Did you initiate a study that involved the testing of treatment, prevention or diagnostic procedures on human subjects?

_____Yes
___X__No

18(B) Did you complete a study that involved the testing of treatment, prevention or diagnostic procedures on human subjects?

_____Yes
___X__No

If "Yes" to either 18(A) or 18(B), items 18(C) – (F) must also be completed. (Do NOT complete 18(C-F) if 18(A) and 18(B) are both "No.")

18(C) How many hospital and health care professionals were involved in the research project?

_____Number of hospital and health care professionals involved in the research project

18(D) How many subjects were included in the study compared to targeted goals?

_____Number of subjects originally targeted to be included in the study
_____Number of subjects enrolled in the study

18(E) How many subjects were enrolled in the study by gender, ethnicity and race?

Gender:
_____Males
_____Females
_____Unknown

Ethnicity:
_____Latinos or Hispanics
_____Not Latinos or Hispanics
_____Unknown

Race:
_____American Indian or Alaska Native
_____Asian
_____Blacks or African American
_____Native Hawaiian or Other Pacific Islander

_____White
_____Other, specify:_____
_____Unknown

18(F) Where was the research study conducted? (List the county where the research study was conducted. If the treatment, prevention and diagnostic tests were offered in more than one county, list all of the counties where the research study was conducted.)

19. Human Embryonic Stem Cell Research. Item 19(A) should be completed for all research projects. If the research project involved human embryonic stem cells, items 19(B) and 19(C) must also be completed.

19(A) Did this project involve, in any capacity, human embryonic stem cells?
_____Yes
___X__ No

19(B) Were these stem cell lines NIH-approved lines that were derived outside of Pennsylvania?
_____Yes
_____ No

19(C) Please describe how this project involved human embryonic stem cells:

20. Articles Submitted to Peer-Reviewed Publications.

20(A) Identify all publications that resulted from the research performed during the funding period and that have been submitted to peer-reviewed publications. Do not list journal abstracts or presentations at professional meetings. **Include only those publications that acknowledge the Pennsylvania Department of Health as a funding source** (as required in the grant agreement). List the title of the journal article, the authors, the name of the peer-reviewed publication, the month and year when it was submitted, and the status of publication (submitted for publication, accepted for publication or published.). Submit an electronic copy of each publication, listed in the table, in a PDF version 5.0.5 format, 1,200 dpi. Filenames for each publication should include the number of the research project, the last name of the PI, the number of the publication and an abbreviated research project title. For example, if you submit two publications for PI Smith for the "Cognition and MRI in Older Adults" research project (Project 1), and two publications for PI Zhang for the "Lung Cancer" research project (Project 3), the filenames should be:

Project 1 – Smith – Publication 1 – Cognition and MRI
Project 1 – Smith – Publication 2 – Cognition and MRI
Project 3 – Zhang – Publication 1 – Lung Cancer
Project 3 – Zhang – Publication 2 – Lung Cancer

If the publication is not available electronically, provide 5 paper copies of the publication.

Note: The grant agreement requires that recipients acknowledge the Pennsylvania Department of Health funding in all publications. Please ensure that all publications listed acknowledge the Department of Health funding. If a publication does not acknowledge the funding from the Commonwealth, do not list the publication.

Title of Journal Article:	Authors:	Name of Peer-reviewed Publication:	Month and Year Submitted:	Publication Status (check appropriate box below):
1. None				☐Submitted ☐Accepted ☐Published

20(B) Based on this project, are you planning to submit articles to peer-reviewed publications in the future?

Yes___X_____ No_____

If yes, please describe your plans:

We plan to submit a manuscript based on our results obtained with the C-07 DASL data.

21. Changes in Outcome, Impact and Effectiveness Attributable to the Research Project. Describe the outcome, impact, and effectiveness of the research project by summarizing its impact on the incidence of disease, death from disease, stage of disease at time of diagnosis, or other relevant measures of outcome, impact or effectiveness of the research project. If there were no changes, insert "None"; do not use "Not applicable." Responses must be single-spaced below, and no smaller than 12-point type. DO NOT DELETE THESE INSTRUCTIONS. There is no limit to the length of your response.

We believe that the major impact for this research project is yet to be realized but should be forthcoming in the next 1 or 2 years when we will have completed the analysis of the C-07 nCounter data and will be able to provide a gene expression signature that will identify patients who are unlikely to receive benefit from oxaliplatin, and, therefore, spare them the toxicities associated with oxaliplatin treatment.

Another important impact our work will have in the next year will be the demonstration that the nCounter platform provides a methodology for gene expression analysis that can easily be adapted to the analysis of RNAs isolated from FFPETs from colon cancer samples collected in clinical trials. The nCounter assays provide many advantages over other clinical expression assays and may spur the development of more cost-effective signatures for a whole range of diseases.

22. Major Discoveries, New Drugs, and New Approaches for Prevention Diagnosis and Treatment. Describe major discoveries, new drugs, and new approaches for prevention, diagnosis and treatment that are attributable to the completed research project. If there were no major discoveries, drugs or approaches, insert "None"; do not use "Not applicable." Responses must be single-spaced below, and no smaller than 12-point type. DO NOT DELETE THESE INSTRUCTIONS. There is no limit to the length of your response.

We provided a new approach for the discovery of prognostic and predictive genes by using the nCounter assays, and we hope to be able to use this technology to improve colon cancer prognosis and treatment decsions.

23. Inventions, Patents and Commercial Development Opportunities.

23(A) Were any inventions, which may be patentable or otherwise protectable under Title 35 of the United States Code, conceived or first actually reduced to practice in the performance of work under this health research grant? Yes_____ No___X___

If "Yes" to 23(A), complete items a – g below for each invention. (Do NOT complete items a - g if 23(A) is "No.")

a. Title of Invention:

b. Name of Inventor(s):

c. Technical Description of Invention (describe nature, purpose, operation and physical, chemical, biological or electrical characteristics of the invention):

d. Was a patent filed for the invention conceived or first actually reduced to practice in the performance of work under this health research grant?
Yes_____ No____

If yes, indicate date patent was filed:

e. Was a patent issued for the invention conceived or first actually reduced to practice in the performance of work under this health research grant?
Yes_____ No____
If yes, indicate number of patent, title and date issued:
Patent number:
Title of patent:
Date issued:

f. Were any licenses granted for the patent obtained as a result of work performed under this health research grant? Yes_____ No____

If yes, how many licenses were granted?_____

g. Were any commercial development activities taken to develop the invention into a commercial product or service for manufacture or sale? Yes___ No____

If yes, describe the commercial development activities:

23(B) Based on the results of this project, are you planning to file for any licenses or patents, or undertake any commercial development opportunities in the future?

Yes__X_____ No_____

If yes, please describe your plans:

It is still unclear whether the models that we are currently building will lead to a model that will improve the prognosis or the prediction for benefit from oxaliplatin. However, should we develop a signature that is validated in the validation cohort, then we may consider licensing or patenting to allow for the possible commercial development of the test.

24. Key Investigator Qualifications. Briefly describe the education, research interests and experience and professional commitments of the Principal Investigator and all other key investigators. In place of narrative you may insert the NIH biosketch form here; however, please limit each biosketch to 1-2 pages. *For Nonformula grants only – include information for only those key investigators whose biosketches were not included in the original grant application.*

Dr. Soonmyoung Paik, M.D.

Dr. Soonmyoung Paik, M.D. is the director of the Division of Pathology at the NSABP Foundation, Inc. and the Director of Samsung Cancer Research Institute. His education in the field of Medicine was at the College of Medicine at Yonsei University, Korea (1975-1981). His Pathology Residency took place at Shadyside Hospital Department of Pathology in Pittsburgh, PA (1982-1985) and at the Department of Pathology at SUNY Health Science Center at Brooklyn, NY (1985-1987).

Dr. Paik's positions and employment are as follows:

1987-1988	Senior Staff Fellow, Medicine Branch, National Cancer Institute, Bethesda, Maryland
1988-1990	Postdoctoral Fellow, Lombardi Cancer Center, Georgetown University, Washington, DC
1990-1995	Assistant Professor of Pathology, Georgetown University, Washington, DC; Director of Tumor Bank Core Facility, Lombardi Cancer Center, Georgetown University (Director: Dr. Marc Lippman)
1995-1999	Co-director, Pathology Section, NSABP Foundation, Inc.
1999-present	Director, Division of Pathology, NSABP Foundation, Inc.
2009-present	Director and Samsung Distinguished Scientist in Medicine, Samsung Cancer Research Institute

Dr. Paik's other experience and professional memberships include:

1999	Steering Committee, Breast Cancer Think Tank, NCI
2000-present	Strategy Group, Program for the Assessment of Clinical Cancer Tests (PACCT), NCI
2001-present	NCI-EORTC Quality Assurance Working Group, Member
2001	NCI Estrogen Receptor Negative Breast Cancer Working Group
2002	NCI PACCT Node Negative Breast Cancer Working Group
2002-present	Research Evaluation Panel, NCI CBCTR (Cooperative Breast Cancer Tissue Resource)
2003-present	ATAC Trial Pathology Committee
2003	NCI Critical Molecular Pathways meeting
2004	NCI Group Banking Committee
2009-present	NCI Breast Cancer Steering Committee
2009	St. Gallen Breast Cancer Symposium, Panelist

Dr. Paik has co-developed and co-authored all NSABP phase 3 clinical protocols developed since 1995. Selections from over 60 of Dr. Paik's peer-reviewed publications are as follows:

1. Paik S, Fisher ER, Fisher B, et al. Pathological findings from NSABP project (protocol B-06) : Prognostic significance of erbB-2 gene expression in breast cancer. J Clinl Oncol 8:103-112, 1991
2. Paik S, Bryant J, Park C, et al. erbB-2 and response to doxorubicin in patients with axillary lymph node-positive, hormone receptor negative breast cancer. J Natl Cancer Inst 90:1361-1370, 1998

3. Paik S, Bryant J, Tan-Chiu E, et al: HER2 and choice of adjuvant chemotherapy for invasive breast cancer: NSABP Protocol B-15. J Natl Cancer Inst 92:1991-8, 2000
4. Paik S, Bryant J, Tan-Chiu E, et al. Real-world performance of HER2 testing--National Surgical Adjuvant Breast and Bowel Project experience. J Natl Cancer Inst 2002 Jun 5;94(11):852-4
5. Paik S: Incorporating Genomics into the Cancer Clinical Trial Process. Seminars in Oncol, 28:305-309, 2001 J Natl Cancer Inst. 2003 Feb 19;95(4):302-7.
6. Paik S, Shak S, Tang G, et al: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer New Engl J Med 2004, 351:2817-2826.
7. Paik S, et al: Technology Insight: application of molecular techniques to formalin fixed paraffin embedded tissues from breast cancer (Review). Nature Clinical Practice Oncology 2:246-254, 2005
8. Carlson RW, Moench SJ, Hammond ME, Perez EA, Burstein HJ, Allred DC, Vogel CL, Goldstein LJ, Somlo G, Gradishar WJ, Hudis CA, Jahanzeb M, Stark A, Wolff AC, Press MF, Winer EP, Paik S, Ljung BM; NCCN HER2 Testing in Breast Cancer Task Force. HER2 testing in breast cancer: NCCN Task Force report and recommendations. J Natl Compr Canc Netw. 2006 Jul;4 Suppl 3:S1-22; quiz S23-4.
9. Paik S, Tang G, Shak S, et al Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol. 2006 Aug 10;24(23):3726-3734.
10. Wolff AC, Hammond ME, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, Dowsett M, Fitzgibbons PL, Hanna WM, Langer A, McShane LM, Paik S, Pegram MD, Perez EA, Press MF, Rhodes A, Sturgeon C, Taube SE, Tubbs R, Vance GH, van de Vijver M, Wheeler TM, Hayes DF.: American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer. J Clin Oncol 25:118-45, 2007
11. Paik, S., Kim, C, Wolmark, N. (2008). HER2 status and benefit from adjuvant trastuzumab in breast cancer. N Engl J Med. Mar 27;358(13):1409-11
12. Ross DT, Kim CY, Tang G, Bohn OL, Beck RA, Ring BZ, Seitz RS, Paik S, Costantino JP, Wolmark N. Chemosensitivity and stratification by a five monoclonal antibody immunohistochemistry test in the NSABP B14 and B20 trials. Clin Cancer Res. 2008 Oct 15;14(20):6602-9. PubMed PMID: 18927301; PubMed Central PMCID: PMC2722068.
13. Kim C, Taniyama Y, Paik S. Gene expression-based prognostic and predictive markers for breast cancer: a primer for practicing pathologists. Arch Pathol Lab Med. 2009 Jun;133(6):855-9. Review. PubMed PMID: 19492877; PubMed Central PMCID: PMC2692069.
14. Goldhirsch A, Ingle JN, Gelber RD, Coates AS, Thürlimann B, Senn HJ; Panel members. Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009. Ann Oncol. 2009 Aug;20(8):1319-29. Epub 2009 Jun 17. PubMed PMID: 19535820; PubMed Central PMCID: PMC2720818.
15. Pusztai L, Jeong JH, Gong Y, Ross JS, Kim C, Paik S, Rouzier R, Andre F, Hortobagyi GN, Wolmark N, Symmans WF. Evaluation of Microtubule-Associated Protein-Tau Expression As a Prognostic and Predictive Marker in the NSABP-B 28 Randomized Clinical Trial. J Clin Oncol. 2009 Aug 10. [Epub ahead of print] PubMed PMID: 19667268.

Dr. Paik's ongoing research support includes:

- NIH/NCI Grant U10 CA12027: National Surgical Adjuvant Breast and Bowel Project
- NIH/NCI Grant U24 CA114731: NSABP Biological Specimen Banks
- NSABP Industry Cores
- Health Research Formula Fund 2009: Discovery and Validation of MicroRNAs as Biomarkers in Breast and Colon Cancer

Dr. Paik's Completed Research Support includes:

- Health Research Formula Fund 2003: Discovery and Validation of Predictors of Response to Anthracycline Based Chemotherapy for Breast Cancer
- Health Research Formula Fund 2004: Comprehensive Survey of Clinical Significance of Amplified Genes in Breast and Colon Cancer
- Health Research Formula Fund 2005: Gene Expression Profiling of BCPT Cancer Events for NSABP Genomic Database, Prediction of Trastuzumab Response in Breast Cancer
- Health Research Formula Fund 2006: Prediction of Response to Bevacizumab in Colon Cancer