# SF 424 R&R and PHS-398 Specific

## Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF 424 R&amp;R Face Page</td>
<td>2</td>
</tr>
<tr>
<td>Research &amp; Related Other Project Information</td>
<td>3</td>
</tr>
<tr>
<td>Abstract</td>
<td>4</td>
</tr>
<tr>
<td>Relevance</td>
<td>5</td>
</tr>
<tr>
<td>Facilities and Other Resources</td>
<td>6</td>
</tr>
<tr>
<td>Equipment</td>
<td>15</td>
</tr>
<tr>
<td>PHS 398 Cover Page Supplement</td>
<td>16</td>
</tr>
<tr>
<td>PHS 398 Research Plan</td>
<td>18</td>
</tr>
<tr>
<td>Specific Aims</td>
<td>19</td>
</tr>
<tr>
<td>Research Strategy</td>
<td>20</td>
</tr>
<tr>
<td>Inclusion Enrollment Report</td>
<td>32</td>
</tr>
<tr>
<td>Human Subjects</td>
<td>33</td>
</tr>
<tr>
<td>Inclusion of Women and Minorities</td>
<td>34</td>
</tr>
<tr>
<td>Targeted/Planned Enrollment Table</td>
<td>35</td>
</tr>
<tr>
<td>Inclusion of Children</td>
<td>36</td>
</tr>
<tr>
<td>Vertebrate Animals</td>
<td>37</td>
</tr>
<tr>
<td>Select Agent Research</td>
<td>38</td>
</tr>
<tr>
<td>Consortium/Contractual Arrangements</td>
<td>39</td>
</tr>
<tr>
<td>Resource Sharing Plan</td>
<td>40</td>
</tr>
<tr>
<td>PI: Newcomb, Polly A</td>
<td>Title: Molecular Correlates of Outcomes in Clinical Trials of Colon Cancer</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>FOA: PA11-260</td>
</tr>
<tr>
<td></td>
<td>FOA Title: RESEARCH PROJECT GRANT (PARENT R01)</td>
</tr>
<tr>
<td>1 RO1 CA176272-01</td>
<td>Organization: FRED HUTCHINSON CAN RES CTR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Senior/Key Personnel:</th>
<th>Organization:</th>
<th>Role Category:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yingye Zheng</td>
<td>Fred Hutchinson Cancer Research Center</td>
<td>Co-Investigator</td>
</tr>
<tr>
<td>Polly Newcomb Ph.D.</td>
<td>Fred Hutchinson Cancer Research Center</td>
<td>PD/PI</td>
</tr>
<tr>
<td>Andrew Chan M.D.</td>
<td>Massachusetts General Hospital</td>
<td>MPI</td>
</tr>
<tr>
<td>Steven Alberts M.D.</td>
<td>Mayo Clinic</td>
<td>Co-Investigator</td>
</tr>
<tr>
<td>Carolyn Hutter Ph.D.</td>
<td>Fred Hutchinson Cancer Research Center</td>
<td>Co-Investigator</td>
</tr>
<tr>
<td>Ulrike Peters Ph.D.</td>
<td>Fred Hutchinson Cancer Research Center</td>
<td>Co-Investigator</td>
</tr>
<tr>
<td>Daniel Sargent Ph.D.</td>
<td>Mayo Clinic</td>
<td>Co-Investigator</td>
</tr>
<tr>
<td>Qian Shi Ph.D.</td>
<td>Mayo Clinic</td>
<td>Co-Investigator</td>
</tr>
<tr>
<td>Stephen Thibodeau Ph.D</td>
<td>Mayo Clinic</td>
<td>Co-Investigator</td>
</tr>
<tr>
<td>Charles Fuchs M.D.</td>
<td>Dana-Farber Cancer Institute</td>
<td>Other (Specify)-Other Significant Contributor</td>
</tr>
<tr>
<td>Howard McLeod</td>
<td>University of North Carolina at Chapel Hill</td>
<td>Other (Specify)-Other Significant Contributor</td>
</tr>
<tr>
<td>Peter Kraft Ph.D.</td>
<td>Presidents and Fellows of Harvard College</td>
<td>Other (Specify)-Other Significant Contributor</td>
</tr>
<tr>
<td>Soonmyung Paik M.D.</td>
<td>National Surgical Adjuvant Breast and Bowel Project</td>
<td>Other (Specify)-Other Significant Contributor</td>
</tr>
</tbody>
</table>
RESEARCH & RELATED OTHER PROJECT INFORMATION

1. * Are Human Subjects Involved?  ☑ Yes  ☐ No
   1.a If YES to Human Subjects
       Is the Project Exempt from Federal regulations?  ☐ Yes  ☑ No
       If no, is the IRB review Pending?  ☑ Yes  ☐ No

2. * Are Vertebrate Animals Used?  ☑ Yes  ☐ No

3. * Is proprietary/privileged information included in the application?  ☑ Yes  ☐ No

4a. * Does this project have an actual or potential impact on the environment?  ☐ Yes  ☑ No

5. * Is the research performance site designated, or eligible to be designated, as a historic place?  ☐ Yes  ☑ No

6. * Does this project involve activities outside of the United States or partnerships with international collaborators?  ☐ Yes  ☑ No
Despite advances in early detection and treatment, colon cancer remains a leading cause of cancer death in the United States. Traditional clinical and pathological features of colon cancer are inadequate in predicting survival. Patient characteristics, such as germline genetic variation, may provide additional prognostic information. To date, only a few studies have used candidate gene and pathway approaches to identify germline genetic factors related to cancer outcomes, and no comprehensive genome-wide studies have been conducted to assess broader genetic variation in relation to outcomes after colon cancer diagnosis. Randomized clinical trials (RCTs) provide a powerful setting in which to address this gap in knowledge, because the patient populations are well-characterized, treatment is standardized, and follow-up is uniform. The overarching goal of the proposed study is to identify genetic factors associated with colon cancer clinical outcomes and to assess whether incorporation of genetic variants improves existing prognostic models. To achieve this goal, we will leverage the resources of three NCI-sponsored phase III RCTs of colon cancer and a recently approved collaboration with the Center for Inherited Disease Research. We propose to use a discovery-based search of host genome-wide single nucleotide polymorphism data in over 6,500 stage II-III colon cancer patients participating in three RCTs, all of whom received 5-fluorouracil / leucovorin / oxaliplatin (FOLFOX) chemotherapy with or without adjuvant therapy, to examine associations with clinical outcomes. Specifically, we will evaluate associations between common genetic variation and disease-free and overall survival among patients with stage II-III colon cancer (Aim 1). We will also evaluate germline genetic variation in relation to treatment-associated serious adverse events (Aim 2). Finally, we will evaluate the impact of adding genetic data into existing web-based, publicly available prognostic models to determine if germline genetic loci can be used, in combination with patient characteristics and clinical factors, to more accurately predict colon cancer outcomes (Aim 3). These prognostic models will be validated in an independent sample of 1,000 stage II-III colon cancer patients from the three RCTs. Data yielded by these investigations will be among the first describing genome-wide germline genetic factors associated with colon cancer prognosis. These results have high translational potential for informing prognosis, augmenting both current strategies that rely on traditional clinical factors and emerging strategies that incorporate information on somatic molecular alterations. Moreover, identifying loci associated with clinical outcomes after colon cancer diagnosis may provide mechanistic insight into cancer progression and metastasis, potentially illuminating new therapeutic targets that can be exploited for clinical translation.
RELEVANCE

This project aims to identify genetic variants that influence clinical outcomes in patients with colon cancer enrolled in clinical trials, and to develop statistical models that integrate evidence on these genetic factors with other known prognostic factors to better predict prognosis. The results of this work may help inform clinical and public health interventions for colon cancer, provide insights into the biology of colon cancer progression, and, ultimately, decrease mortality associated with this disease.
The unique scientific environment and accumulated resources of the Fred Hutchinson Cancer Research Center support the proposed project. Formed in 1973, the Hutchinson Center is one of the world’s leading cancer research centers. Home of three Nobel Laureates, the Hutchinson Center is an independent, nonprofit research institution dedicated to the development and advancement of biomedical research. Its organizational mission is the elimination of cancer, HIV/AIDS, and other diseases as causes of human suffering and death. In 2002, the Hutchinson Center took the lead role in forming the Fred Hutchinson/University of Washington Cancer Consortium, a research collaboration comprising Fred Hutchinson Cancer Research Center, the University of Washington (UW), Seattle Children’s, and the Seattle Cancer Care Alliance (SCCA). While the participating institutions remain independent, the Consortium combines their strengths to accelerate research progress in the fight against cancer and build premier research programs across the disciplines of basic, clinical and public health sciences. The Consortium is one of 41 National Cancer Institute-designated comprehensive cancer research centers nationwide. In addition, the Hutchinson Center enjoys productive industry collaborations that enhance its ability to achieve advances in cancer research and treatment.

The Hutchinson Center is organized into the four scientific divisions of Basic Sciences, Clinical Research, Public Health Sciences (PHS), and Human Biology; and the divisions of Administration and Development. The Center’s Vaccine and Infectious Disease Institute (VIDI) was created in 2007 to develop and implement prevention strategies for globally important infectious diseases, including HIV, malaria, and the viruses and bacteria that can result in cancer. VIDI is made up of three disciplines: population sciences, infectious disease sciences, and immunology and vaccine development; VIDI faculty working in the population sciences are PHS faculty members.

**Division of Public Health Sciences (PHS)**

Research in the PHS Division studies cancer in populations, focusing on determining causes of cancer, helping to identify and assess effective screening and treatment methods, developing prevention strategies that reduce the risk of cancer, and developing research strategies to assist people to change behaviors toward healthier lifestyles. This prevention-oriented research also extends to other diseases, including HIV/AIDS, cardiovascular disease, diabetes, and fractures. The specific experience and expertise of the Division includes (1) the conduct of research in methodology for the design and analysis of rigorous follow-up studies, (2) the provision of statistical services needed for ensuring well-designed and properly conducted and analyzed clinical trials, (3) the evaluation of measures intended to reduce the incidence and mortality from cancer and other chronic diseases; to develop and test screening and surveillance methods that can be readily translated into clinical practice; to assess the outcomes of these maneuvers; and to train students and post-doctoral fellows in prevention research methods through formal grant-supported training programs. Close relationships to the UW and to specialized clinics at the SCCA have enhanced research and service opportunities of the Cancer Prevention Program.

**The Cancer Prevention Program (CPP)** at the Hutchinson Center is the oldest such program in the nation, and one of the largest in the world, with more than a million people around the globe participating in studies aimed at reducing cancer incidence and mortality and improving quality of life. Many of these studies are also relevant to the control of other chronic diseases. The CPP consists of 47 member-track faculty (regular and joint) investigators, 13 PhD staff scientists, and 31 affiliate faculty from diverse disciplines, and 297 staff members. CPP faculty actively participate in training the next generation of scientists; they currently supervise 41 students and post-doctoral trainees. The 2010 research base totaled more than $27 million in direct funding, with 180 continuing grants and contracts. Cancer Prevention continues its interdisciplinary goals to address research questions that relate to cancer etiology and risk factors, including the identification of genetic, biologic, nutritional, and behavioral characteristics; to conduct clinic and community-based intervention studies in targeted populations in order to identify ways to reduce morbidity and mortality from cancer and other chronic diseases; to develop and test screening and surveillance methods that can be readily translated into clinical practice; to assess the outcomes of these maneuvers; and to train students and post-doctoral fellows in prevention research methods through formal grant-supported training programs. Close relationships to the UW and to specialized clinics at the SCCA have enhanced research and service opportunities of the Cancer Prevention Program.
Laboratory
The Specimen Processing Shared Resources provides a broad range of services including the development and validation of procedures for specimen processing analyses; processing of blood, urine, buccal and DNA samples and DNA extraction. Resource operations are split into three sites, for a total of 1,451 net square feet. Space is provided on the C level of the Hutchinson building, 5th floor of the Arnold Building at the Hutchinson Center's South Lake Union Campus, and a satellite lab is located on the 7th floor of the Seattle Care Alliance. All necessary facilities and instrumentation required for processing and tissue culture activities are provided.

Clinical
N/A - Clinical facilities will not be required for the conduct of this research.

Animal
N/A - Animal facilities will not be required for the conduct of this research. The FHCRC maintains a fully accredited animal care facility for those projects engaged in animal studies.

Computing Systems and Environment
FHCRC Information Technology Services (Center IT). FHCRC provides support for a complex heterogeneous computing environment. The network infrastructure consists of multiple Local Area Networks (LANs) connected to the Fred Hutchinson Network. Intel personal computers, Macintosh systems, and LINUX/UNIX workstations are connected to servers running different operating systems. The FHCRC network is connected to the Internet through the Pacific Northwest GigaPop Network (PNWGP) with 1 Gbit/s. The FHCRC Network is protected by redundant firewalls and a redundant intrusion prevention system. There are over 3,000 local workstations, printers, and servers connected to the FHCRC Network. FHCRC supports both UNIX/IMAP and Microsoft Exchange for e-mail services. Fax and photocopier machines are also available throughout the Center. The FHCRC computing environment serves the needs of the proposed project well. Not only does it provide a mechanism for investigators to collaborate with colleagues within the project and within the Center, but also through the PNWGP Network with Consortium partners and through the Internet with institutions around the world. FHCRC IT manages the FHCRC Storage Area Network (SAN) which consists of a 3PAR storage backend and is sharing data via a NetApp storage gateway to allow maximum collaboration of Windows, Linux and Macintosh computer users. The total networked storage capacity is currently 350TeraByte. The current system is expected to grow to 1.2 PetaByte by 2012. Data protection is implemented by DataDomain appliances in conjunction with Commvault Simpana backup software and IBM Tivoli Storage Manager.

FHCRC also provides support for over 60 applications that are made available to the entire Center.

PHS Divisional Computing Shared Resources. The Division provides services that are specific to programs under the PHS umbrella that would otherwise be duplicated by each group. These include computer-training labs, divisional-shared presentation equipment, and development servers that are made available to projects for prototyping and testing of new applications (apart from production networks). LINUX statistical applications are also available from Divisional Computing Shared Resources server farm clusters. The Computing Shared Resource has staff with project installation/management, database and application design, and training design experience available for consultation. Specifically there are two large computing resources available for PHS researchers:

(1) The Hyrax cluster is currently equipped with 250 compute nodes (1940 cores / 5.3 TB RAM) and connected to a high performance NetApp storage array. It consists of the following:

49 Intel Xeon nodes, each with 8 cores/8 GB RAM for a total of 392 cores/392 GB RAM.

63 AMD Opteron nodes, each with 4 cores/8GB RAM for a total of 252 cores/504 GB RAM.

90 AMD “Shanghai” nodes, each with 8 cores/32 GB RAM for a total of 720 cores/2880 GB RAM.

48 AMD “Istanbul” nodes, each with 12 cores/32 GB RAM for a total of 576 cores/1536 GB RAM.

Each cluster node possesses a local disk drive which can be used for non-shared temporary data. All of the cluster nodes are 64-bit Linux systems running the current OpenSuSE distribution. A variety of standard software is installed on the head node and on each work node. This includes the current version of R which is updated on a monthly basis as well as the OpenMPI and MPICH parallel processing frameworks. The overall cluster is managed by a software package named SLURM, the Simple Linux Utility for Resource Management. SLURM controls the assignment and scheduling of all jobs running on the cluster.
The Orca cluster is currently equipped with 8 large memory nodes which offer 64-128 GB RAM memory. These systems are mainly used for prototyping and preprocessing of large datasets such as GWAS (genome wide association studies) data and proteomics.

**PHS Divisional Network Support** The network allows PC's, Macintosh, and LINUX workstations continuous 24-hour access to resources. All physical servers have redundant power and network connections, and use RAID 5 to insure protection against drive failure. In order to accommodate the growing demand for computer resources and mitigate the power demands of physical systems, many of the server services are now being provided by virtual systems using the advanced VMWare ESX technology. These virtual systems are supported by multiple hosts and have a high degree of redundancy. Physical servers are backed-up daily through Center-supported tape backup system. The virtual systems have the added protection of being recoverable through snapshots which are taken and stored on a daily basis.

The main PC operating system is Microsoft Windows XP Professional. The primary application suite used on all computers is Microsoft Office 2003 Professional Enterprise Edition. Remote users can access data servers through the Internet by using a secured Virtual Private Network tunnel. Statistical software used in the program includes R, SAS, SPSS, STATA, SPLUS and MATLAB. Database applications supported in the program include Microsoft SQL Server, Oracle, Sybase, and Access.

Special survey research support for computer-assisted telephone interviewing, automated open-ended coding, data entry software and optical scanning is available from the Consortium's Collaborative Data Services shared resource. The CDS Creative Services unit provides support in the preparation of high quality materials utilizing a variety of graphics packages on both PC and Macintosh platforms. HP Color LaserJet and Epson Stylus Pro printers are available for high quality and large format printing production.

**Security**
The Center has a tipping point Intrusions prevention system and a clustered Nokia firewall.

**Other**
Special survey research support for computer-assisted telephone interviewing, automated open-ended coding, data entry software and optical scanning is available from the Consortium's Collaborative Data Services shared resource. The CDS Creative Services unit provides support in the preparation of high quality materials utilizing a variety of graphics packages on both PC and Macintosh platforms. HP Color LaserJet and Epson Stylus Pro printers are available for high quality and large format printing production.

The Comprehensive Center for the Advancement of Scientific Strategies (COMPASS) supports research aimed at eliminating cancer and other diseases of public health importance. COMPASS offers a strong foundation for health research studies by providing study coordination to scientific investigators. The COMPASS organization gained experience establishing and coordinating the Carotene and Retinol Efficacy Trial (CARET), a multi-center lung cancer prevention trial that began in 1983. Since that time, COMPASS has expanded and now manages several large-scale studies. They provide both Project Management and Informatics & Data Management including: Information Technology and Informatics (Web site design and maintenance; Electronic communication services; Data sharing standards and methods; Computer and network systems development and support); and Database Development and Management (Quality assurance systems; Tracking system development; Reporting; Data collection and processing; Database design and maintenance).

**Office Facilities**
Office space and furnishings are available within the Cancer Prevention Program, which is allocated 50,955 net square feet in the Robert M. Arnold PHS Building, a state-of-the-art facility located on the Center's Robert W. Day Campus. Core program facilities available are located on Levels D and 1-5, and include offices for scientists, research associates and research support staff. In addition, research projects within the PHS Division can make use of specialized spaces for conferences, telephone and in-person interviewing, exercise studies, feeding studies and medical examination functions. The on-line Event Management System allows for scheduling appropriately-sized conference rooms and support services, including individualized room set-up, computerized projection, video conference connections, and catering for meetings held within the Arnold Building and across the Day Campus. PHS researchers have easy access to the Consortium shared resources and other Hutchinson Center scientific divisions and facilities, as well as increased interactions with colleagues in PHS Laboratories (located on Level 5 of the Arnold Building).

**Arnold Library Shared Resource**
The Arnold Library is a state-of-the-art, well-integrated library service at the Hutchinson Center with the primary purpose of providing high quality, responsive and up-to-date services in support of the Center's research, education and patient care programs. It is one of the essential components of the information infrastructure that supports and facilitates biomedical inquiry and scientific progress for investigators at the Hutchinson Center. This centralized resource provides
access to journals (print collections and online databases), monographs, bibliographic information, and extensive electronic resources in basic and public health sciences and biomedical research. Services range from intra-library loan/document delivery to assisting authors by depositing their papers in Pub Med Central, in compliance with NIH Public Access Policy.

Massachusetts General Hospital

Scientific Environment
The facilities and resources of the Massachusetts General Hospital (MGH) are uniquely suited to conduct this investigation of molecular correlates of outcomes of clinical trials of colon cancer. The investigators at MGH have specific expertise in the maintenance, conduct, and analysis of clinical trial cohorts which comprise the study population. The proposed study benefits from the unique features of the scientific environment which includes leading experts in chronic disease epidemiology, biostatistics, genomics, and prediction modeling. Moreover, the environment is able to recruit talented trainees and post-doctoral fellows that can lead analyses and learn from senior investigators.

Office
The MGH Gastrointestinal Unit (Dr. Chan) has several thousand square feet of research and administrative space, including a clinical research unit located adjacent to the hospital campus. Dr. Chan will coordinate the study, meet with study personal, perform data analysis, and prepare manuscripts in his 300 square foot office. All postdoctoral researchers and research assistants have their own desks in shared offices or in open space.

Laboratory
The Chan laboratory at MGH consists of 600 square feet of space including 4 workstations for computational and biostatistical analyses. There is ready access to SAS statistical packages, including an interface to the Channing Laboratory and Harvard School of Public Health computer systems which for data storage, including GWAS data.

Computer
The MGH Gastrointestinal Unit has extensive state-of-the-art computer facilities with access to a full complement of UNIX & Windows NT servers & workstations that form the heart of a sophisticated network of Macintosh and PC computers and printers. In addition, study personnel will have access to a UNIX system with SAS programming capabilities through the Channing Laboratory, which maintains all data for Harvard cohort studies, including NHS, HPFS, and PHS. The Channing Laboratory computer facility provides access to Linux and Solaris based servers for centralized authentication, email, security, audit, automated backup and recovery, web servers, data storage and analysis. The Channing computer system consists of thirty servers and several hundred PCs networked for primary data management and analysis operations. The servers range from a Sun Enterprise 3000 with 2 gigabytes of memory to Sun Fire X4240s with 32 gigabytes of memory. They share 40 terabytes of disk storage, tape drives, and printers. Available software includes SAS, R, Splus, SUDAAN, Proc StatXact, Stata, Matlab and gauss, extensive custom software used for management and analysis of large complex datasets and an extensive list of bioinformatics software. An Oracle database management system is used to manage all biospecimen data. Available software includes SAS, ISML, Splus, and extensive custom software programs needed for management and analysis of large, complex data sets.

Core Facilities and Other Resources
All MGH investigators will have access to Countway Library, one of the largest medical and health libraries in the United States. Countway Library contains over 2 million books and receives 4,000 journal titles. They have public access to Medline and the best search service in the country. Additional statistical support for this proposal is also provided through the Harvard Catalyst Program and the MGH Clinical Research Program.

Clinical
N/A - Clinical facilities will not be required for the conduct of this research.

Animal
N/A - Animal facilities will not be required for the conduct of this research.

Mayo Clinic
Mayo provides an ideal environment in which to provide training to the next generation of cancer clinical trials and patient reported outcomes methodological and collaborative researchers. This environment includes a diverse and dedicated faculty with depth of representation in relevant disciplines, a substantial research base affording numerous opportunities for trainees to become integrated in ongoing collaborative research projects, high quality support facilities that include the
Mayo Clinic Cancer Center's core laboratories. Equally important, but more challenging to describe, is the scientific climate at Mayo, in which there are no barriers to collaboration across clinical, basic, and observational disciplines. A description of key institutional facilities, resources and support processes that are used extensively by Center members and administration follows.

Research Environment
Mayo Medical Center in Rochester, Minnesota, occupies approximately 14 million square feet of space. There are two hospitals—Saint Marys and Methodist. The rest of the medical center space is utilized for: i.) patient evaluation, radiologic and laboratory diagnostic services; ii.) medical center administration; iii.) a variety of medical and research-based graduate educational programs; and, iv.) biomedical research. Total space available for all research is approximately 728,000 square feet. Research programs at Mayo reflect the interest and initiative of individual investigators. There are approximately 177 established research programs. Nearly 350 research fellows are pursuing advanced fulltime research training at Mayo. The NIH supports research at Mayo, with more than 329 MD and PhD investigators supported by approximately 690 active NIH grants. Various other extramural funding sources, including the Department of Defense Medical Research Initiative, private foundations, and industry, also support research at Mayo Clinic Rochester.

Department of Health Sciences Research
The Department of Health Sciences Research (HSR) at Mayo Clinic is a multidisciplinary group of more than 50 doctoral level and 400 allied health staff who are dedicated to improving patient care through medical research. The department has a rich heritage dating back to Dr. Joseph Berkson (Chair 1933-1964) and Dr. Leonard Kurland (Chair 1964-1987). In 1935, Dr. Berkson developed a computerized medical diagnostic coding system, capturing all patient diagnoses and procedures in order to efficiently facilitate clinical research. Building on Dr. Berkson's work, in 1966 Dr. Kurland initiated the Rochester Epidemiology Project, which has become a national resource for population-based epidemiological studies with continuous funding from the NIH for over 40 years resulting in more than 1500 peer-reviewed publications to date. Current faculty and staff also conduct leading edge methods research on a wide range of areas, such as bioinformatics, natural language processing, medical terminology standards, genetic analysis methods, analysis of DNA microarray data, proteomics, quality of life measures, molecular epidemiology, and health economics and utilization. The Department of Health Sciences Research is located primarily in the Harwick Building across the street from the Mayo and Gonda Buildings, the main outpatient facilities. The department consists of three divisions (Epidemiology; Biomedical Informatics and Biostatistics; and Health Care Policy and Research), all of which play major roles in clinical and translational research at Mayo Clinic. Below are descriptions of the two divisions primarily involved in the design, implementation and analysis of the proposed training program as well as a description of the computing facilities within the department.

Division of Biomedical Statistics and Informatics
The Biomedical Statistics and Informatics personnel occupy over 10,000 square feet on the 11th floor of the Stabile building and over 26,000 square feet on the 7th and 8th floors of the Harwick Building. The Division has provided consultation on statistical design and analysis for the clinical and laboratory research staff at Mayo since this service was introduced by Joseph Berkson (of "Berkson's bias" fame) in 1932. There are currently 26 Ph.D. statisticians, and 55 statisticians with Master's degrees, whose activities are facilitated by 77 statisticians with Bachelor's degrees ("statistical programmer analysts"), and 8 clerical personnel. The Division also has 10 PhD trained bio- and medical informaticists, supported by 25 IT personnel. In addition to general consulting on over 2000 ongoing investigations, this group provides core statistical support for a number of program project grants as well as for the Mayo Clinic Cancer Center and the Rochester Epidemiology Project. The Division also maintains an active research effort, especially in the areas of epidemiologic modeling, survival analysis, statistical genetics and in the design, early stopping and analysis of clinical trials, and patient reported outcomes. Within Biostatistics, a major effort has been made during the past few years to improve standardization and greater efficiency of resources. There are central repositories for genetic SAS macros, Unix scripts, and R/Splus functions, all stored with revision control and agreed-upon programming standards. Documentation for over 100 software packages, written locally and downloaded from elsewhere, is easily accessible from the web. Examples of basic analysis plans are being documented as well as 'how-to' guides indicating how the software can best accomplish these plans. Experts are available for a variety of study designs, including those that are family based and non-family based. In addition to collaboration with investigators from across the institution, multiple members of the Division including faculty members on the proposed training program are actively engaged in cutting edge statistical research. The primary data analysis tools in the Division are SAS® and S-PLUS®. These are augmented with other tools such as StatXact, Data Conversion, R, WinBugs, and Matlab. The Division has been a user of SAS® since its 1972 release and has made significant contributions to the SAS® software base, including many of the survival routines. Members of the Division have authored all of the survival analysis routines distributed in the S-Plus package, including routines to handle mixed effects Cox models. The locally written S-Plus recursive partitioning and genetic libraries are also publicly available. Routines developed for internal use include 4 SAS® procedures, 7 S-Plus® libraries each consisting of, numerous functions, and over 250 SAS® macros, including a special macro developed for estimating incidence and
prevalence in the Olmsted County population. Programs with this level of distribution require extreme dedication to accuracy, testing, and reliability with the goal of establishing easy-to-use, reliable, and leading edge analysis tools.

The Division has two principal goals related to bioinformatics: i.) the indexing and organization of clinical data generated by the clinical practice for search and retrieval in support of research and education; and, ii.) a program of basic research on clinical concept representation, information indexing, and database retrieval. Presently, more than 40 members, including faculty, programmers, and support personnel are involved in biomedical informatics related work. This group traces its origin to the advent of a structured medical record at Mayo Clinic in 1907 and the retrieval indices based on tabulating equipment, and subsequently computers, that has made Mayo an unparalleled resource for clinical research. During the past five years, the group has established a close working relationship with IBM Life Sciences Division to design, prototype, and deploy a comprehensive information archive of basic science research observations, genomic data, patient clinical data, natural language processed documents, and metadata, to comprise the largest single repository of detailed clinical and research observations associated with medical record manifestations, the Mayo Life Sciences System. Over the past 20 years, the biomedical informatics group has sustained an international reputation for basic and applied informatics research, predominantly in the area of biomedical concept representation and ontologies. Additionally, it has developed world-class expertise in medical natural language processing (NLP), data integration, retrieval query interpretation, and information transforms. More recently, the Division has assumed a major role in caBIG, the Biomedical Informatics Grid program sponsored by the NCI among comprehensive cancer centers throughout the United States. Members of the division have played leadership roles at HL7 (the dominant health information standards effort internationally), ISO TC215 on Health Informatics, the ANSI Health Informatics Standards Board, and numerous related academic forums.

Research Resources
There are extensive research resources and facilities at Mayo that can be brought to bear on the training of fellows. These are described briefly, below:

The Research Environment at Mayo. The clinical research environment at Mayo is extremely productive with 1,800 to 2,000 protocols reviewed at the Mayo Institutional Review Board yearly and approximately 4,000 peer-reviewed reports published by Mayo authors annually.

Mayo Clinical Practice. Mayo Clinic is one of the largest and most respected healthcare institutions in the world, providing multidisciplinary integrated services for approximately 400,000 patients each year. Demand for care at Mayo is increasing and as a result the institution is currently undergoing a dramatic expansion, both in structure and of personnel. Patient care, education and research at Mayo are based on true teamwork, with the expectation of collaboration and excellence. Administrative separations between clinical disciplines and between basic and clinical investigators, which can be rate limiting at some academic health centers, are minimized at Mayo. Instead, there is a commitment to programmatic initiatives, such as the Cancer Center, to facilitate translation of intellectual and technologic expertise.

North Central Cancer Treatment Group (NCCTG). The NCCTG is an NCI-sponsored, cooperative group of 40 community clinics that enters approximately 2,500 patients per year on Phase II and Phase III clinical trials in breast cancer, lung cancer, gastrointestinal cancers, neuro-oncology, hematology, and cancer prevention and control. Mayo serves as the research base for the NCCTG, providing Mayo investigators with a unique resource with which to test clinical and translational research questions. In particular, with the recent six-year renewal of the NCCTG grant in 2006, monies were allocated to permit the collection of DNA samples from patients enrolled in randomized controlled therapeutic trials. This resource complements the long-standing practice of collecting and maintaining tissue blocks of trial participants.

Mayo Clinic Cancer Center (MCCC). An NCI-designated comprehensive cancer center, MCCC has had continuous funding from NCI for over 30 years. The Center is vital and growing: The Center has implemented a wide-ranging Education Program that includes broad services for patients and the public, in addition to its professional course offerings. As of the 2008 grant renewal, there are a total of 12 programs, including 7 new programs and 5 previously approved programs. The MCCC is noted for high-quality clinical research, clinical trials, multidisciplinary translational working teams and biostatistics. The QOL core component has been rated as outstanding during a recent competitive renewal. The Center has a robust patient base with approximately 4,000 new patients seen in Medical Oncology and 5,000 in Hematology annually. Patients seek care at Mayo for common cancers and for a spectrum of uncommon cancer-related conditions. These patients have been enthusiastic participants in medical research studies, and it is not uncommon to enroll over 90% of eligible subjects in our observational research studies.

Biostatistics Shared Resource, of the MCCC has provided expert statistical collaboration with MCCC investigators for over 30 years. The purpose of the resource is to provide statistical and biomathematical collaboration for the development and conduct of peer-reviewed cancer research generated by investigators in the MCCC. This research ranges from the bench to the bedside to the population, and encompasses basic science, clinical trials, epidemiologic research, and other
translational and educational research. The primary usage of the funds provided to the resource is to support statistical and mathematical collaboration on pilot projects, for assistance to investigators in developing approved research projects not otherwise funded but leading towards external funding, and to support unanticipated needs for statistical collaboration on MCC approved projects.

The Biostatistics Shared Resource has multiple focus areas (teams) tied together into a well-organized, efficient core. These are: i.) a Clinical Trials team responsible for cancer clinical trials, associated translational research, interventional psychosocial research, and imaging, together with patient and public education research projects; ii.) a Population Science team responsible for providing statistical collaboration and data management support for cancer observational studies, including genetic and molecular epidemiology; iii.) a Quality of Life team responsible for providing general and specific advice, measurement tools, and analysis for MCC investigators who are evaluating the affect of clinical and psychosocial interventions on cancer patients, families, caregivers, and others; and iv.) a Biomathematics team that provides collaborative expertise in the mathematical aspects of imaging, data processing and analysis, mathematical modeling (e.g. of tumor growth), and general mathematical or algorithmic issues. These groups work efficiently and effectively together to provide statistical and biomathematical collaboration to investigators.

The Mayo Clinic Survey Research Facility (SRF), provides investigators with support for the conduct of surveys. This support includes expert consultation, instrument development, sampling, mailing, telephoning, response tracking, and instrument processing. Having conducted over 1,000 survey-based studies since its beginning in 1991, the SRF has experience in many areas pertinent to cancer research (e.g., risk factor assessment, obtaining outcomes following diagnosis or treatment, recruitment for clinical trials, assessing attitudes and behaviors related to cancer screening, assessing the acceptability of cancer screening methods, determination of family trees for genetic studies, assessing satisfaction with cancer related treatment and associated health care encounters, assessing informational needs for cancer patients, assessment of the willingness to help others to stop smoking (cancer prevention), psychological scale development pertinent to intervening on adolescents who use tobacco, and quality of life assessment in cancer patients).

Mayo Core Facilities
A major strength supporting research at Mayo is a set of excellent Core Laboratories. The Core Facilities are available to all investigators and students, and provide the following resources: centralized use of state-of-the-art instrumentation; theoretical and technical expertise; collaborative interactions, both with core personnel and with other investigators who may use similar techniques; training for students or other personnel who wish to learn techniques in detail, as well as the theoretical basis of various techniques. Each of the cores is headed by an accomplished investigator and is staffed by highly skilled technical professionals. These resources allow faculty and students to follow their research along virtually any path of interest, and with minimal concern of technological limitations.

The Advanced Genomics Technology Center — Eric Wibben, PhD, Director. The Advanced Genomics Technology Center (AGTC) brings together Genomics-oriented shared resources to facilitate interactions with the Mayo research staff and to increase the coordination and efficiency of our activities. With the support and cooperation of the Genomics Research Center, the Comprehensive Cancer Center, and the CTSA, the facilities of the AGTC are working to provide professional, efficient, and low-cost access to the latest Genomics technologies to all Mayo investigators. The AGTC also operates laboratory-based training programs in basic Molecular Biology and Tissue Culture techniques. Together, these facilities occupy about 20,000 sq. ft. of laboratory space. The directors of the facilities meet on a regular basis to improve coordination and workflow, and to discuss opportunities for improving our services for the community. The resources in the AGTC include: i.) Cytogenetics — Routine cytogenetics plus SKY, FISH and CGH analysis; ii.)DNA Sequencing — Full service automated DNA sequence analysis; iii.) DNA Synthesis — Custom synthesis of oligonucleotide probes, primers, RNAs and molecular beacons; iv.) Genotyping — Linkage analysis (microsatellites), SNP analysis, genome-wide screens; v.) Microarray — Transcriptional profiling using Affymetrix microarrays and custom spotted arrays; vi.) Biospecimens Accessioning and Processing (BAP) — Specimen handling, processing, and preservation plus nucleic acid extraction; and vii) Tissue and Cell Molecular Analysis (TACMA) — Immunohistochemistry, in situ hybridization, and digital image analysis.

RNA Interference Core Facility: The AGTC recently announced the opening of a new core facility, which provides lentiviral shRNA knockdown vectors to all Mayo investigators. Access to the newly established RNA Interference Technology Resource (RITR) is available through the AGTC web page. The RITR interface provides a 'shopping cart' approach to searching for and purchasing Sigma-Aldrich RNAi lentiviral shRNA knockdown vectors, which, as a result of our partnership with Sigma-Aldrich, are available to Mayo investigators at a substantially reduced price. The AGTC has done extensive testing with these knockdown vectors. Eleven investigators in Rochester and Jacksonville were involved in the beta testing phase, and over 40 different human and mouse genes have been successfully knocked down.

Immunohistochemistry Core Laboratory — George G. Klee, MD, PhD, Joseph McConnell, PhD, and Ravindar J. Singh, PhD, Co-Directors. The Immunohistochemistry Core Laboratory (ICL) provides laboratory testing at minimal cost to Mayo researchers. When capacity is available, testing is also provided for investigators outside Mayo on a collaborative basis. The ICL staff is actively involved in new assay development and improvement of current assay methodology. Requests for ICL service are prioritized based on volume, number of protocols requiring the test(s), availability of an alternative source.
for testing and development effort required. First priority for laboratory testing is given to CTSA protocols and other NIH funded protocols.

**Computer**

Mayo has extensive computer resources. Those of particular relevance to the proposed training program include: The Health Sciences Research departmental system, the Central Computer Facility, and the Research Computer Facility (RCF). All computer systems are connected to a comprehensive and centrally administered network consisting of multiple fiber optic (FDDI) rings that span the campus. These are connected via Cisco routers to area or building backbones (FDDI and/or Ethernet) which are in turn connected to individual floor devices via UTP concentrators.

The Department of Health Sciences Research has a suite of approximately thirty Sun Unix servers and several Linux servers which comprise the main computing platform of the department. Roughly half of the Sun servers support a general purpose computer structure with a full unix desktop. These servers are typically Sun V440's with 4 1.6GHz Spare III processors and 8G of memory. Various applications are run including but not limited to SAS, Splus, R, plus custom or other off the shelf application code. A departmental file server supports more than 1.2Tb of "home drive" and "projects" space that can be mounted under Unix or Windows. Additionally, a Sun Grid Engine with access to 7 Sun T2000 "Coolthreads" servers for a total of 160 threads/processors. This Grid Engine has access to over 8T of high speed "scratch" disk for large data analysis. This Grid Engine is utilized 24/7 and runs thousands of jobs every month. Additional servers include: Sybase database for general purpose relational database applications; a specialized Ingres database supporting Mayo's Cancer Center; a filenet image server currently has over 10 million pages of archived medical records used for retrospective studies; a departmental web server which serves to present up-to-the minute results to internal investigator as well as more general purpose web content; and several department web servers with dedicated engines for tomcat, SaslntrNet, and custom built daemons. Most personnel have desktop Windows XP or Linux workstations connected by 100MB or 1GB links to the network. Network connected, shared high-volume printers are in close proximity to all personnel with color and large-paper (11x17) devices available on all floors. The HSR Citrix environment provides a full Windows desktop to Statisticians and other HSR staff who use Linux workstations. We currently support about 185 concurrent users with 300 total users. Citrix provides productivity applications such as the Microsoft Office suite, e-mail access and Internet browsing to our users. It also gives them access to the Windows versions of a number of statistical programs like SAS, S-Plus, R and Winbugs. Users are able to access the same data in both Citrix and Linux as well as cut and paste between both environments. The HSR Citrix farm consists of 15 Dell 1750 and 1950 servers. They are all dual processor Intel servers with 4GB of memory and 76GB hard disks. The servers run the 32 bit version of Windows 2003 server with Citrix Presentation 4.0 and are protected by Trend anti-virus software.

The Mayo Research Computing Facility (RCF) has provided computational infrastructure support for research at Mayo since 1986. The majority of the RCF staff members and resources are located in the Rosa Parks Pavilion Building on the Rochester downtown campus, although some of our staff are located within the areas they support. The goal of the Research Computing Facility is to guide, coordinate, and enable the effective use of computing and information management technologies by the Mayo investigator. Resources available consist of a staff of computing professionals well versed in addressing the needs and interests of research using hardware and software tailored to the needs of the investigator. The RCF currently offers three integrated services that support investigators conducting research in a heterogeneous computing environment: i.) **Systems Infrastructure Support** integrates and supports the various "hidden" technologies used in Mayo's research labs, providing cross-platform access to file servers, databases, shared printers, the Internet, server-based software, and research-specific projects and resources. The Systems team also provides data storage, backup, and recovery services, giving Mayo research more than ten terabytes of storage space. In addition, the RCF administers and maintains a Beowulf-style Linux cluster, a high-powered compute engine, and several file servers. These systems give research areas access to a number of bioinformatics tools, including Bioperl, BLAST, EMBoss, HMMER, PHYLIP, and PolyPhred; ii.) The **Collaborative Consulting Resource** supports the objectives of individual research programs, emphasizing technology integration and providing custom application development. The CCR also tries to identify areas of common need among research departments, which leads the CCR to design common, reusable solutions for the benefit of all RCF clients; and, iii.) **RCF Education** offers training on an as-needed basis for non-standard computers and software used in Mayo's research areas. In addition to customized classes and training opportunities, the RCF also provides on-line and printed materials for learning essential computer skills and programs that are not taught in standard Mayo classes.

The Laboratory Information System contains records on every patient sample processed by Mayo Medical Laboratories. The Central Computer Facility maintains key system-wide data bases including the registration system (containing demographic data), billing and resource utilization, and the emerging Mayo electronic medical record. A long-term archiving facility retains important Mayo information in perpetuity.

**Office**

Each trainee will be provided office space equipped with a computer and administrative assistant support.
Other Resources
The Mayo Clinic Library System was established in 1907 and includes a central clinical and research library in Rochester, Minnesota, and a number of other specialized libraries in Rochester and the Mayo Group Practice sites in Jacksonville and Scottsdale. The collections cover the entirety of clinical medicine, biomedical research areas, and allied fields. Mayo library collections include approximately 350,000 volumes and subscriptions to 4,300 scientific, technical, and medical periodicals. The Mayo libraries are full-service libraries supporting a broad range of information products and services. Mayo libraries are fully automated utilizing the Innovative Interfaces, Inc. product for acquisitions, cataloging, serials control, reserves, media booking, and a system-wide on-line catalog. A bibliographic retrieval system supporting 10 databases (including MEDLINE, Current Contents, EMBASE) called MayoSearch is based on the Ovid Technologies, Inc. product. A full-text drug information system with the Micromedex suite of databases and systems as the primary content is also available. A rapidly-growing list of full-text journals, full-text newswire alerts, and a consumer health information database are also accessible at workstations throughout the Mayo campus. Library web sites provide access to electronic resources on the Mayo intranet and serve as launch sites for the various library electronic reference systems such as the MayoSearch bibliographic retrieval system, Foundation on-line catalog, and other specialized services.

The Section of Publications has editors, proofreaders, editorial assistants and other staff to assist with the preparation of scientific papers.

Audiovisual, Video Communications and Medical Graphics/Visual Information Sections with trained full-time staff are available for the development and preparation of audiovisual instruction material, including television tapes.
EQUIPMENT

Fred Hutchinson Cancer Research Center

Standard office equipment (photocopiers, facsimile machines, paper shredders, etc.) will be available to all project investigators and staff within Cancer Prevention. All offices are equipped with new Cisco Voice over IP (VoIP) telephones with voicemail. The Center's department of Voice/Data Operations provides common voice and data services for the Center, including backbone local area networks, a wide area network, Internet connectivity, telephones, voice mail, remote access, and firewall.

Massachusetts General Hospital

Standard office equipment (photocopiers, facsimile machines, printers, telephones, etc.) will be available to all participating faculty and staff at Massachusetts General Hospital. All the necessary computing equipment, network resources, and internet access will also be available.

Mayo Clinic

Standard office equipment (photocopiers, facsimile machines, printers, telephones, etc.) will be available to all participating faculty and staff at the Mayo Clinic. All the necessary computing equipment, network resources, and internet access will also be available.
### 1. Project Director / Principal Investigator (PD/PI)

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Mr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>* First Name</td>
<td>Polly</td>
</tr>
<tr>
<td>Middle Name</td>
<td>A.</td>
</tr>
<tr>
<td>* Last Name</td>
<td>Newcomb</td>
</tr>
<tr>
<td>Suffix</td>
<td>Ph.D.</td>
</tr>
</tbody>
</table>

### 2. Human Subjects

- Clinical Trial? [x] No [ ] Yes
- * Agency-Defined Phase III Clinical Trial? [ ] No [ ] Yes

### 3. Applicant Organization Contact

Person to be contacted on matters involving this application

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Mr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>* First Name</td>
<td>Mark</td>
</tr>
<tr>
<td>Middle Name</td>
<td>A.</td>
</tr>
<tr>
<td>* Last Name</td>
<td>Boyer</td>
</tr>
<tr>
<td>Suffix</td>
<td></td>
</tr>
</tbody>
</table>

* Phone Number: (206) 667-4868  Fax Number: (206) 667-6221
* Email: gmail@fhcrc.org

* Title: Director, Office of Sponsored Research
* Street 1: Fred Hutchinson Cancer Research Center
* Street 2: PO Box 19024, J6-500
* City: Seattle
* County/Parish: King
* State: WA: Washington
* Country: USA: UNITED STATES  *Zip/Postal Code: 981091024
4. Human Embryonic Stem Cells

* Does the proposed project involve human embryonic stem cells?  
  ☒ No ☐ Yes
1. **Application Type:**

From SF 424 (R&R) Cover Page. The response provided on that page, regarding the type of application being submitted, is repeated for your reference, as you attach the appropriate sections of the Research Plan.

*Type of Application:

- [x] New
- [ ] Resubmission
- [ ] Renewal
- [ ] Continuation
- [ ] Revision
SPECIFIC AIMS

Despite recent declines in colon cancer mortality, the disease remains a leading cause of cancer-related death in the United States. Overall, estimated 5-year survival for colon cancer is 65%; however, survival rates vary widely. Stage at diagnosis is currently the strongest predictor of outcome. Approximately 20% of colon cancer patients present with stage IV disease, or distant metastases, and experience uniformly dismal survival. Prognosis is more favorable for those with non-metastatic disease, but less than 65% of patients with stage II-III colon cancer survive to 5 years post-diagnosis without recurrence. Surgery with "curative intent" remains the primary treatment modality for these cases, and adjuvant chemotherapy with 5-fluorouracil, leucovorin, and oxaliplatin (FOLFOX) has also a routine treatment approach for stage II-III disease. However, clinico-pathological staging is inadequate in distinguishing between stage II-III patients who would be cured by surgery alone and those at high risk for disease recurrence. Considerable efforts are underway to develop novel prognostic markers for colon cancer, with a focus on the role of somatic genomic variants in colon tumors. In contrast, relatively little attention has been paid to the role of variation in the host genome as a determinant of colon cancer prognosis. Recent studies that have examined host genetic variation in relation to clinical outcomes in patients with other cancers have yielded highly significant results. As noted in a recent Nature Genetics editorial, "...an additional lesson [of these studies] for cancer genomics is the relevance of both the host and the tumor genomes in understanding cancer progression and treatment response."

The overarching goal of this study is to identify common germline genetic variants associated with colon cancer survival outcomes and treatment-associated serious adverse events, via a discovery-based search of the host genome.

To achieve this objective, we propose to leverage genome-wide single nucleotide polymorphism (SNP) data from 6,557 patients with stage II-III colon cancer who received standardized regimens of FOLFOX adjuvant chemotherapy when enrolled in NCI-sponsored phase III randomized clinical trials (RCTs). Colon cancer prognosis is known to vary, even within stage, according to patient characteristics and clinical factors such as age, sex, tumor location, and histological grade. Thus, the key innovation in our approach is the examination of host genetic factors within a population of colon cancer patients in which these factors are well-characterized, treatment is standardized, and follow-up for disease outcomes and treatment-associated serious adverse events is uniformly conducted. To clinically translate our results, we will integrate knowledge gained with respect to prognostic host genetic factors into existing, widely used prognostic models to determine if such genetic factors can be used in combination with traditional information on patient characteristics and clinical factors to more accurately predict colon cancer clinical outcomes. We propose the following specific aims:

Aim 1. Identify germline genetic loci associated with survival outcomes among patients with stage II-III colon cancer. We hypothesize that host genetic variation is associated with disease-free and overall survival.

Aim 2. Identify germline genetic loci associated with treatment-associated serious adverse events among patients with stage II-III colon cancer. We hypothesize that host genetic variation is associated with treatment-associated serious adverse events, such as gastrointestinal toxicities, neutropenia, and paresthesias.

Aim 3. Examine the impact of adding information on germline genetic loci to existing prognostic models for stage II-III colon cancer currently based on patient characteristics and clinical factors, such as age, sex, tumor location, tumor stage, and histological grade. We hypothesize that the addition of host genetic variation data will improve the performance of these models in predicting prognosis among patients who receive standard adjuvant chemotherapy (FOLFOX).

The proposed study is cost-effective and timely. Our highly experienced, transdisciplinary team of investigators will efficiently pool resources from well-characterized extant study populations to pursue a comprehensive assessment of germline genetic variation in relation to colon cancer outcomes, while accounting for clinical factors. Using a discovery-based approach to uniformly search the genome in an unbiased manner facilitates the investigation of several pathways, and is comparable in cost and DNA requirements to a candidate gene approach. Data yielded by these investigations will be among the first characterizing host genetic factors associated with colon cancer prognosis, and will be highly complementary to ongoing efforts in these RCTs to further clarify the prognostic impact of intratumoral molecular markers and clinicopathologic factors. Results will have high translational potential, augmenting both current approaches that rely on traditional clinical factors and emerging strategies that incorporate intratumoral molecular alterations as prognostic markers. Moreover, identifying loci associated with outcomes may provide mechanistic insight into cancer progression and metastases, potentially suggesting new therapeutic targets that can be exploited for clinical translation.
RESEARCH STRATEGY

A. SIGNIFICANCE
In the United States alone, over 101,000 men and women are diagnosed with colon cancer and more than 50,000 die from the disease each year. Approximately 20% are initially diagnosed with stage IV, or metastatic disease, which is associated with uniformly poor survival. For the remaining 80% of colon cancer patients, surgery with "curative intent" remains the primary treatment; however, current American Joint Committee on Cancer staging fails to optimally identify those patients who can be cured by surgery alone. As a result, adjuvant chemotherapy with 5-fluorouracil, leucovorin, and oxaliplatin (FOLFOX) has become routinely recommended for stage III patients and some stage II patients. However, the majority of stage II-III patients do not benefit such adjuvant chemotherapy: 33% are cured by surgery alone, and 25% develop recurrence despite this adjuvant treatment. Thus, identification of novel prognostic factors that can better distinguish colon cancer patients at high versus low risk for cancer recurrence and other serious clinical outcomes is a clear unmet need for both individual patients and the broader health care system.

Presently, the few known prognostic factors for colon cancer beyond stage are primarily clinicopathologic features, including nodal status, histological grade, and age. Tumor markers, such as microsatellite instability status and somatic KRAS mutations, have been associated with prognosis and may be predictive of response to certain cancer therapies; however, these markers are not yet routinely used to assess prognosis or guide treatment in clinical practice. Although several groups, including a team led by Dr. Sargent (Co-I, Mayo), are actively researching the clinical utility of intratumoral prognostic markers, relatively little attention has been paid to the potential prognostic significance of germline genetic markers. Information on the role of germline genetic factors with respect to cancer progression and treatment response represents an important gap in knowledge that could add to our ability to predict prognosis and deliver treatment in a tailored manner.

To date, linkage and genome-wide association studies (GWAS) have successfully identified several genetic variants associated with risk of incident colon cancer. In contrast, comparatively little is known regarding host genetic factors associated with colon cancer outcomes in patients with established disease. Recently, we examined GWAS-identified susceptibility loci in relation to mortality in 2,611 colorectal cancer (CRC) patients and found a significant association between only one locus, a SMAD7 gene polymorphism, and survival. Although promising, our work, and that of others, indicates that the common genetic variants that influence colon cancer survival are likely different from those that influence risk. A necessary next step is a more comprehensive analysis encompassing a broader assessment of germline genetic variation through genome-wide scans of single nucleotide polymorphisms (SNPs) in relation to clinical outcomes in colon cancer patients.

Importantly, genome-wide scans for survival outcomes among patients with extant colon cancer have a strong likelihood of influencing clinical practice. Compelling recent examples of the clinical utility of such findings include identification of a polymorphism in TCL1A in the IL-17A pathway associated with musculoskeletal adverse events among breast cancer patients enrolled in a randomized clinical trial (RCT) of aromatase inhibitors, and identification of an association between a variant in IL-17F and two-fold poorer overall survival in a phase III RCT of advanced pancreatic cancer. Such findings have direct clinical implications, and may be utilized in a "personalized" approach to identify patient groups at high risk of poor clinical outcomes.

An additional key aspect of our proposed work is the incorporation of genetic markers into two highly utilized prognostic models developed by our group and others: Numeracy and Adjuvant!. Presently, these two models include information on patient and clinical characteristics to predict prognosis. Although both models are well-validated, publicly available, and widely used in clinical practice, their performance is not optimal, and additional predictors are needed to further improve their clinical utility. With the advent of expanded options for cancer treatment, most of which incur substantial toxicity and cost, it is imperative to develop improved prognostic models that can enhance individualized therapeutic decision-making.

Taken together, this cost efficient and timely project will provide an unprecedented opportunity to examine the relationship between germline genetic variation and prognosis, providing both mechanistic insights and clinical translation. The knowledge gained has the potential to be transformative, leading to improvements in clinical practice and advancing our understanding of pathways for colon cancer treatment.

B. INNOVATION
Our proposal is conceptually innovative in its examination of the contribution of common germline genetic variation to clinical outcomes (Aims 1 and 2). First, most studies to date investigating novel prognostic markers in colon cancer have focused on intratumoral markers, including somatic mutations, genomic alterations, or gene
We propose to significantly augment this approach by comprehensively examining germline genetic variation in relation to clinical outcomes in colon cancer patients. Second, to our knowledge, this will be the first large-scale genome-wide examination of common genetic variation and colon cancer outcomes. The few prior studies of germline genetic factors and colon cancer outcomes have used a candidate gene approach and have often been limited by small sample sizes, which have contributed to inconsistent findings. In contrast, we will employ a broad, unbiased genome-wide approach with a large sample size and validation to identify SNPs associated with clinical outcomes. Third, existing prognostic models, although well-validated and utilized by many in clinical practice, are based on a limited number of clinical and patient features and do not optimally distinguish high- and low-risk subgroups. We will incorporate of germline genetic variants into these models, which may improve assessments of prognosis and better identify high- and low-risk subgroups for tailored cancer treatment and surveillance.

This proposal is methodologically innovative in its leverage of resources from recent NCI-sponsored phase III colon cancer RCT patient populations for the examination of host genetics in relation to clinical outcomes. We will combine the resources of the only such completed RCTs of non-metastatic colon cancer that have collected specimens for the isolation of germline DNA suitable for genome-wide SNP genotyping. In the context of these well-characterized studies with standardized treatment, we will be able to carefully account for other key prognostic factors. The standardized, rigorous monitoring and reporting for disease recurrence and treatment-associated serious adverse events will allow us to evaluate the relationship between host genetics and a number of outcomes. We will efficiently use data and biospecimens previously collected in these RCTs, and will apply these materials to evaluate important research questions complementary to but beyond the scope of the original RCT protocols. Moreover, the resources generated herein will be highly complementary to other ongoing funded and planned studies within these RCTs examining additional prognostic markers, including intratumoral molecular features. Together, these efforts will form the basis for future work examining genomic predictors of colon cancer molecular subtypes, and the combined influence of germline and somatic genetic alterations outcomes.

C. APPROACH

C.1 Overview of Study Design

The overall goal of this study is to identify germline genetic loci associated with colon cancer outcomes. An overview of our approach is provided in Figure 1. Using data from three RCTs of stage II-III colon cancer, we propose to identify common germline genetic variants associated with survival outcomes (Aim 1) and treatment-associated serious adverse events (Aim 2). We will also use information on SNPs associated with survival outcomes to build upon existing, validated prognostic models (Aim 3). Our primary outcomes include disease-free survival (DFS) and overall survival (OS) for Aims 1 and 3, and treatment-associated serious adverse events (SAES) for Aim 2. This proposal leverages existing, well-annotated RCT data, as well as new genome-wide scan data not routinely available in the RCT setting. Genotyping for two of the RCTs will be provided through an approved application through the Center for Inherited Diseases Research (CIDR), contingent on receipt of funding through the present proposal (letter, Dr. Camilla Day). The third RCT will have been genotyped under separate funding (letters, Drs. Charles Fuchs and Howard McLeod).

C.2 Specific Aim 1

AIM 1: Identify germline genetic loci associated with survival outcomes among patients with stage II-III colon cancer. We hypothesize that host genetic variation is associated with disease-free and overall survival.

C.2.a. Rationale

To date, at least 20 variants in 17 loci associated with colorectal cancer (CRC) susceptibility have been identified through GWAS. Nonetheless, recent work has suggested that as many as 65-70 susceptibility loci remain to be identified, accounting for -17% of the total genetic variability in CRC susceptibility. Most SNPs identified through GWAS were not targets of earlier candidate gene studies, underscoring a key advantage of the unbiased GWAS approach to SNP discovery.
Despite the success of GWAS for CRC susceptibility, to date there has been no comprehensive genome-wide examinations of SNP data for survival in patients with CRC or, more specifically, colon cancer. However, there is increasing evidence that colon cancer survival has a genetic component. In a secondary analysis of one phase III RCT for stage III colon cancer, family history of colon cancer was associated with a 28% improvement in DFS.52 Two *All point estimates adjusted for sex and age and stage at diagnosis. studies in Sweden have also showed familial aggregation in CRC prognosis among CRC cases with a parent who also had CRC, whose those parent died from CRC 510 years after diagnosis had a 1.44-fold [95% confidence interval (CI): 1.01-2.01] greater risk of dying from CRC themselves,53 and, more generally, the risk of death from CRC for people with a parent who died from CRC was 1.20-fold (95% CI: 1.03-1.41) higher than that in people with a parent who had the disease but died from other causes.54 To date, most studies of common genetic variation in relation to colon cancer or CRC survival have used candidate approaches, focusing on putative pathways associated with cancer therapies,25 mismatch repair genes,61 or candidate genes such as COX2,62 NAT2,63 TLR-3,64 and SMAD7.65 Limited by small sample sizes, these studies have reported null to marginally significant associations with minimal validation. Four recent studies, including our own, examined loci identified from GWAS for CRC susceptibility in relation to survival. In a study of 7 variants,25 a significantly lower risk of death was associated with the rs4779584 SNP on 15q13 [hazard ratio (HR)=0.33, 95% CI: 0.15-0.72]. In contrast, a separate study evaluating 10 variants,26 and another focused on variants in the 8q24 region,24 failed to identify any associations with survival. Taken together with our results (see Preliminary Studies below), these studies are consistent with the likely possibility that genetic variants most predictive of colon cancer or CRC survival are different from those associated with cancer risk. A necessary next step is a comprehensive analysis of genome-wide SNP data in relation to survival among patients with colon cancer.

**Preliminary Studies.** Our multi-institutional research team is led by Dr. Newcomb at the Fred Hutchinson Cancer Research Center (FHCRC) and Dr. Chan at Massachusetts General Hospital (MGH), both of whom have substantial expertise in colon cancer genetic and survival analyses. To complement their expertise, our team also includes key leaders of the RCTs in this proposal as co-investigators [Drs. Alberts, Sargent, and Thibodeau (N0147)] and other significant contributors (OSCs) [Dr. Paik (C-08) and Dr. Fuchs (80702)]. These collaborations demonstrate the commitment of the clinical trial cooperative groups to this application.

Recently, Drs. Newcomb and Chan led an analysis in 2,611 CRC cases from five prospective cohort studies evaluating the relationship between survival after CRC diagnosis and 16 SNPs identified from prior GWAS for CRC risk.66 Findings from that analysis suggest that variation in the rs4939827 SNP (SMAD7) is associated with survival (Figure 2), providing proof-of-principle that common genetic variation can provide prognostic information beyond traditional prognostic factors (e.g., stage). However, other evaluated loci were not associated with survival, suggesting that the common genetic variants most informative of disease survival likely differ from those most informative of risk and supporting the need for discovery-based interrogations of the genome focused on survival. These findings illustrate our ability to utilize genomic data in studies of cancer survival, and demonstrate the effectiveness of our investigator team in synthesizing data from multiple cohorts.

In addition to our research detailed above, Dr. Newcomb’s research with the Colon Cancer Family Registry has recently demonstrated 20% lower disease-specific survival in CRC cases who reported pre-diagnostic use of non-steroidal anti-inflammatory drugs relative to never-users,68 with effect modification by polymorphisms in rs4939827 (SMAD7).69 Results from this study also indicate no association between polymorphisms in NFkB1 or pre-diagnostic hormone therapy USE and CRC survival, but note poorer survival in cases who were underweight or obese or who were smokers at the time of diagnosis.74 Dr. Chan’s research with the Nurses’ Health Study and Health Professionals’ Follow-up Study has demonstrated that regular aspirin use after diagnosis is associated with better CRC survival.76 Dr. Chan is also actively investigating associations between CRC tumor characteristics and survival1,33,76-83 and is involved with other correlative studies within NCI-sponsored phase III RCTs. The success of these projects illustrates our experience working with clinical and molecular correlates of colon cancer survival in cohort and RCT populations.

Our research team also has considerable expertise with analyses of genome-wide SNP data.50,55-93 Dr. Peters (Co-I) and Dr. Kraft (OSC) are leaders in this field. Drs. Peters and Huffer (Co-Is), and colleagues, recently published a meta-analysis of GWAS data (2,906 CRC cases / 3,416 controls) with follow-up (8,161 cases / 9,101 controls) from the
**C.2.b. Methodological approach**

We propose a genome-wide scan of SNP data in 5,057 stage II-III colon cancer patients enrolled in two RCTs (N0147 and C-08). Following genotyping, thorough quality control (QC) procedures, and imputation, we will conduct analyses to identify variants that differ according to DFS or OS. Based on analyses in these two RCTs, we will select the most promising variants to be followed up in another RCT (80702).

**Study Populations.** This project will utilize data and peripheral blood specimens collected as part of N0147 and C-08 for initial SNP selection, as well as data from 80702 for follow-up of identified significant SNPs. The design of each of these studies is shown in Figure 3 and described below. In all studies, explicit consent was obtained from participants for use of their blood specimens and data in future genomic research.

**NCCTG N0147.** N0147 is a phase III RCT of FOLFOX with or without cetuximab for treatment of stage III colon cancer lead by the North Central Cancer Treatment Group (NCCTG). Based on findings in the metastatic setting that benefits from cetuximab may be limited to patients with KRAS-wildtype disease, the original study protocol was modified after enrollment began to limit future randomization to KRAS-wildtype patients. Primary results from N0147 were recently reported: among patients with KRAS-wildtype disease, 3-year DFS and OS were not significantly different in patients who received FOLFOX alone (74.6% and 87.3%, respectively) and those who received FOLFOX + cetuximab (71.5% and 85.6%). Germline DNA has been extracted from collected blood specimens for 3,157 patients. In April 2011, we received approval from the NCCTG Concept Review and Prioritization Committee and the NCI Cancer Therapy Evaluations Program (CTEP) GI Intergroup Correlative Sciences Committee to access these biospecimens for genome-wide SNP platform genotyping (letters, Drs. Steven Alberts, Axel Grothey, Margaret Mooney).

**NSABP C-08.** C-08 is a phase III RCT of 2,710 patients with stage II (25%) and III (75%) resected colon cancer lead by the National Surgical Adjuvant Breast and Bowel Project (NSABP). Primary results from C-08 were recently published, indicating no difference in 3-year DFS for patients randomized to receive treatment with FOLFOX with vs. without bevacizumab (77.4% vs. 75.5%, respectively). An estimated 2,400 patients in C-08 provided blood specimens for DNA extraction. DNA will be extracted and shipped, using a 3rd party anonymization procedure in Year 1 of the study. In May 2011 we received approval from the NSABP Specimen Use and Review Committee to access C-08 biospecimens for DNA extraction and genotyping (letters, Drs. Norman Wolmark and Soonmyung Paik).

**CALGB 80702.** 80702 is an on-going phase III RCT for stage III colon cancer, lead by the Cancer and Leukemia Group B (CALGB) research group, that will allow for independent follow-up of SNPs identified in discovery analyses. CALGB 80702 will compare FOLFOX given for 3 vs. 6 months with or without the addition of oral celecoxib (400mg/d). Enrollment began in June 2010 and the study has been accruing as expected. Enrollment is expected to close in 2013 (target n=2,500). We conservatively estimate that 2,000 subjects from this study will have available genome-wide SNP data, collected through the existing study protocol, before Year 4 of this proposal (letters, Drs. Charles Fuchs (OSC) and Howard McLeod (OSC)).
Genotype Data. In September 2011, we received approval for genotyping of NO147 and C-08 on the Illumina OmniExpress platform to be fully funded by CIDR, contingent on funding for the present proposal (91X01HG0066201). This genotyping introduces efficiency by facilitating centralized, rapid genotyping. Genetic data will be centrally processed using a standardized QC pipeline that our FHCRC team has used for the past three years, and will be imputed to the HapMap CEU population using MACH. As part of this QC, we will exclude SNPs based on call rate (<98%), minor allele frequency (<5%), Hardy-Weinberg Equilibrium P-value (<0.0001), and imputation quality r² metric (<0.7). Data will be securely transferred to the Mayo Clinic for analysis in collaboration with the FHCRC and MGH. Genome-wide scans of 80702 will be conducted under separate funding through Dr. Howard McLeod (OSC).

Statistical Analysis. We will consider two outcomes in Aim 1 analyses: DFS and OS. DFS is defined as the time from study randomization to disease recurrence or death from any cause. OS is defined as time from randomization to death from any cause. Based on outcome data as assessed through the most extended available follow-up at the time of analysis, we will use Cox proportional hazards to calculate summary HRs for associations between genome-wide SNP data and survival outcomes, modeling a one degree of freedom (1 df) trend test with the observed or imputed dosage of the minor allele as the independent variable. HRs will be adjusted for age at diagnosis, year of diagnosis, sex, treatment arm, and principal components for population substructure. Reflecting the composition of the patients recruited to these RCTs, 88% of our combined population self-reported White race. We will conduct sensitivity analyses restricted to non-Hispanic Whites, and will perform exploratory analysis for other racial/ethnic groups. We plan to follow-up promising findings in minority populations through future proposals with existing collaborators (see Future Plans).

Discovery, SNP selection, and follow-up joint analysis. We will initially identify SNPs in a discovery phase by analyzing genome-wide SNP data from 5,057 subjects in NO147 and C-08, with regression analyses performed separately in each study. A random sample of 250 cases from each study will be excluded from Aim 1 to be reserved as an independent validation set for prognostic model analyses (Aim 3). We will combine study-specific results via fixed-effects meta-analysis to select SNPs associated with DFS or OS. We will examine I² and 0-statistics, excluding SNPs with strong evidence for heterogeneity between the two studies. We will use Q-Q plots to evaluate over-dispersion of test statistics, applying genomic-control methods to further adjust our statistics with a lambda correction factor if needed.

Based on meta-analyses of NO147 and C-08, we will select the most promising SNPs for follow-up in 80702, using the genome-wide scan data that will be available for this latter study. We will select SNPs with a P-value ≤5.0x10⁻⁵ for follow-up analyses. Assuming that 2 million SNPs pass our QC, as we have observed in GWAS for CRC risk, we expect -10,000 SNPs to pass this P-value cutoff for each of the two evaluated outcomes. In separate analyses, we will evaluate these SNPs in relation to DFS and OS in a random sample of 1,500 participants from 80702; the remaining 500 participants from 80702 will be excluded from Aim 1 to be reserved for independent validation of prognostic models in Aim 3. We will then perform a joint analysis, using fixed-effects meta-analysis to pool the results from discovery (N0147, C-08) and follow-up populations (80702). In these joint analyses we will consider a SNP to be significantly associated with DFS or OS if it attains genome-wide significance (P 5 5x10⁻⁵).

Exploratory analyses. In addition to these primary analyses, we will examine associations between SNPs and survival outcomes according to two novel adjuvant therapeutic agents: cetuximab (N0147) and bevacizumab (C-08) (all treatment arms received FOLFOX). Such analyses will provide an opportunity for identifying SNPs with differential or concordant treatment effects. These analyses will be performed separately in each study, in light of the different treatment regimens used.

Statistical power. As shown in Table 1, assuming OS of 70% and DFS of 65% by the time of analysis, we are well-powered to identify associations with SNPs that have a minor allele frequency (MAF) ≥10%. For MAF=10%, we can detect an HR of 1.30 or greater for the discovery phase and 1.48 in the follow-up phase for OS, with similar power for DFS. Minimum detectable HRs are more favorable with greater MAF. The magnitude of these HRs is substantially more sensitive than those observed in a recent genome-wide survival analysis of a phase III RCT of pancreatic cancer (CALGB 80303) in which a SNP in IL-17F was associated with a greater than two-fold shorter OS (P=2.61x10⁻⁵) in a population of only 294 patients. Our much larger study is thus likely well-positioned to identify SNPs relevant to colon cancer prognosis.

<table>
<thead>
<tr>
<th>TABLE 1. Statistical power for Aim 1*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MAF</strong></td>
</tr>
<tr>
<td><strong>OS</strong> (30%)</td>
</tr>
<tr>
<td>Disc.</td>
</tr>
<tr>
<td>0.50</td>
</tr>
<tr>
<td>0.25</td>
</tr>
<tr>
<td>0.10</td>
</tr>
</tbody>
</table>

Abbreviations: MAF, minor allele frequency; HR, hazard ratio; OS, overall survival; DFS, disease-free survival, Disc., discovery phase; FUP, follow-up phase

*alpha= 5x10⁻⁵ in discovery analyses (N=5,057) and 5x10⁻⁸ in joint follow-up analyses (N=6,557). Power calculations assume a log-additive mode of inheritance.
C.2.c. Expected Outcomes of Aim 1

With the completion of Aim 1, we expect to have identified common genetic variants that are associated with survival after colon cancer diagnosis in patients receiving standardized treatment in RCTs. The large sample size and efficient design will provide results for our subsequent evaluation of the impact of germline genetic loci on existing prognostic models (Aim 3), and will help inform future functional and translational follow-up studies.

C.2.d. Potential Problems and Alternative Strategies

**Heterogeneity within and between studies, particularly with respect to treatment.** The three RCTs included in the present proposal imposed different inclusion criteria and were designed to test distinct treatment interventions. Overall results from these RCTs have indicated no primary treatment effects; however, this does not preclude an effect of specific treatments on SNP associations. To account for this, we will adjust for treatment, and will assess possible treatment interactions with SNPs. We will also explore subgroups by race/ethnicity. We are committed to generating the necessary data for pooled analyses with future RCTs to facilitate larger, more detailed analyses for minority populations through our existing collaborations. **Survivor bias.** Survival is generally more favorable in RCT populations as inclusion is contingent on surviving long enough to be enrolled and the absence of certain co-morbidities. To address these potential concerns, we will account for left truncation and will conduct a series of exploratory analyses to evaluate the robustness of our results. In addition, based on our ongoing work with population-based survival studies, we plan to examine the associations identified in this proposal in more broadly representative patient populations. **Rare variants.** Based on our sample size, we are focusing on common variants (MAF 5%). We will explore imputation to other platforms (e.g. 1000 genomes), which provide more data and coverage, particularly for rare variants. We anticipate beginning genotyping in Year 1, by which time exome chip results in studies of colon cancer will be available. If such results have their expected positive contributions to detecting important disease loci, we will pursue using the newly developed HumanOmniExpress Plus Exome rather than the OmniExpress platform, which should be feasible given the minimal cost differential. This will allow us to leverage additional variant content of potential prognostic significance.

C.3 Specific Aim 2

**AIM 2: Identify germline genetic loci associated with treatment-associated serious adverse events** among patients with stage II-III colon cancer. We hypothesize that host genetic variation is associated with treatment-associated serious adverse events, such as gastrointestinal toxicities, neutropenia, and paresthesias.

C.3.a. Rationale

The nature and degree of treatment-associated serious adverse events (SAEs) experienced by colon cancer patients is highly variable. There is substantial evidence that genetic factors are important in tolerance of standard treatment regimens for colon cancer. Studies to date have shown that genetic variation in specific pathways that are known targets of treatments such as 5-fluorouracil (5-FU) are important in the development of organ system toxicities in colon cancer patients. Specifically, small studies of colon cancer patients treated with 5-FU identified polymorphisms in **TYMS**, the gene responsible for production of thymidylate synthase, as being associated with grade 3 SAEs, including neutropenia and diarrhea. Additionally, genetic variation in the gene that produces methylenetetrahydrofolate reductase (**MTHFR**) appears to be associated with the incidence of SAEs after 5-FU treatment in colon cancer patients with non-metastatic disease. As evidenced by these findings, the approach to identifying genetic loci associated with treatment toxicity has, to date, primarily used a candidate approach to evaluate genetic regions upstream of drug metabolism genes, therapeutic targets, or DNA repair genes. The use of a candidate approach is limited by our understanding of such pathways and of the full spectrum of the physiological effects of treatment; as a result, studies using a candidate approach have been limited in the types of genes analyzed, and have not take into account the complexity of drug combinations. A more comprehensive method is necessary to more fully characterize the complex relationship between genetic variation and SAEs in colon cancer patients.

Recent GWAS for incidence of CRC and other common complex diseases have been successful in identifying loci that were not recognized using a candidate or pathway approach. We thus expect that a genome-wide scan designed to identify SNPs associated with treatment-associated SAEs would similarly identify new genetic variants beyond those in previous candidate studies. Improving our understanding of the genetic basis underlying tolerance to standard treatment regimens for colon cancer has the potential to inform more personalized therapeutic decisions (e.g., via treatment modification or specific monitoring). The results of this study are expected to represent a significant advance in fulfilling the promise of using predictive genomic factors for colon cancer treatment toxicity.

**Preliminary Studies.** Our research group has previously described SAEs in the included RCTs, demonstrating familiarity with their definition, measurement, and interpretation. Dr. Alberts (Co-I) and colleagues recently reported on SAEs in N0147: grade 3 SAEs were recorded for 50.4% (FOLFOX) and 71.2% (FOLFOX+cetuximab) of KRAS-wildtype patients, with a similar magnitude of differences in the KRAS-mutated treatment arms (Table 2). In another
RCT of stage IV CRC (N9741), Drs. McLeod (OSC), Sargent (Co-I), and Thibodeau (Co-I) demonstrated that neutropenia after treatment with irinotecan/oxaliplatin was associated with the homozygous UGT1A1*28 allele (55% vs. 15%; P=0.002); risk of neutropenia after treatment with 5-FU/oxaliplatin was associated with GSTM1 deletions (28% vs. 16%; P=0.02). Additionally, in an adenoma clinical trial, Dr. Chan (M-PI) showed that patients taking celecoxib, the adjuvant used in CALGB 80702, did not appear to experience differential treatment-associated cardiovascular, renal, or gastrointestinal toxicity according to CYP2C9*2 variant, despite the association of this variant with impaired celecoxib metabolism. As with Aim 1, the considerable genetic epidemiology and statistical expertise of our research team is complemented by the clinical and RCT experience of Dr. Alberts, and by Dr. Sargent's statistical experience with the included RCTs. This work demonstrates our ability to successfully analyze genome-wide data of treatment-associated SAEs.

C.3.b. Methodological Approach
To identify germline genetic loci associated with treatment-associated SAEs we will analyze genome-wide SNP data from the two RCTs (N0147 and C-08) as described in Aim 1. However, in Aim 2 we will use this data to identify variants that differ in distribution according to the occurrence of SAEs, both overall and in relation to specific classes of common toxicities. Although the adjuvant therapies evaluated in these RCTs (i.e., cetuximab, bevacizumab) have not been shown to be associated with DFS or OS in stage II-III patients, these therapies do appear to be associated with certain SAEs. Thus, as described below, we will modify our statistical approach to allow for treatment effects and gene-by-treatment interactions in identifying potential genetic variants associated with SAEs. As with Aim 1, we will identify SNPs associated with clinical outcomes using data from N0147 and C-08, and will follow-up the most promising variants in another RCT (80702).

Classification of Adverse Events. Detailed information on SAEs was collected in all RCTs using the Common Toxicity Criteria, version 3.0. Routine reporting of SAEs was required for all events. Grade was calculated as the maximum severity across all cycles of treatment. For this analysis, treatment-associated SAEs will be defined as an overall sum of grade ≥3 events, but we will also conduct analyses evaluating associations with specific classes of more common grade ≥3 SAEs, including gastrointestinal (e.g., diarrhea, vomiting), neutropenia, and paresthesias.

Statistical Analysis. General analytic approaches proposed for Aim 2 parallel those for Aim 1. We will consider an overall summary of SAEs and classes of specific common SAEs as outcomes. We will employ Cox proportional hazards regression models, adjusted for age at diagnosis, year of diagnosis, sex, treatment arm, and principal components for population substructure. As in Aim 1, we will conduct sensitivity analyses restricted to non-Hispanic Whites, with exploratory analysis for other racial/ethnic groups. Instead of using a 1 df test of the minor allele as the independent variable, we will use a 2 df test simultaneously testing the genotype (G) and the genotype by treatment (GxT) interaction. This 2 df test leverages potential treatment and GxT effects, and tests the null hypothesis that the genetic marker is not associated with SAEs in any stratum defined by treatment. This approach has been successfully used in genome-wide settings to identify variants associated with Parkinson's Disease and glycemic traits.

Discovery, SNP selection, and follow-up joint analysis. As with Aim 1, discovery will be done by analyzing SNP data from 5,057 subjects in N0147 and C-08, with regression analyses performed separately in each study. Study-specific results will be combined using Fisher’s method for combining P-values. We will use Q-Q plots to evaluate over-dispersion of test statistics, applying genomic-control lambda correction factor if needed.

We will use results from meta-analyses of N0147 and C-08 to select SNPs for follow-up in 80702 based on a P-value of 5.0x10⁻³, with separate SNP selection for SAEs overall and for specific classes of SAEs (gastrointestinal SAEs, neuropathy (including paresthesias), and neutropenia). As noted in Aim 1, we expect -10,000 SNPs to pass this cutoff.
in each outcome-specific analysis. We will evaluate SNPs identified in discovery analyses in relation to SAEs in 1,500 participants from 80702 using a 2 df Cox proportional hazards regression. We will perform a joint analysis to pool the results from discovery (N01047 and C-08) and follow-up (80702) RCTs. In these joint analyses we will consider a SNP to be significantly associated with the risk of SAEs if it reaches genome-wide significance ($P < 5 \times 10^{-8}$).

**Exploratory analysis.** Although we are taking treatment and GxT interactions into account, the focus of our main analysis will be on FOLFOX-associated SAEs. All treatment arms in each of the included RCTs received FOLFOX, as this is the current standard-of-care for colon cancer in the adjuvant setting. Nonetheless, it will also be important to examine adjuvant-specific SAEs. Therefore, in an exploratory manner we will examine common genetic variation and SAEs for the novel adjuvant therapeutic agents evaluated in NO147 and C-08 (i.e., cetuximab, bevacizumab), as well as for study participants randomized to FOLFOX alone.

**Statistical power.** As shown in Table 3, assuming overall SAE occurrence of 50%, we are well-powered to identify SNPs with an MAF 10% and HR of 1.29 for the discovery phase and 1.45 in the joint follow-up phase. For SAEs with a prevalence of 30% (e.g. grade 3 neutropenia), we are well powered to identify SNPs with a MAF 10% with a HR of 1.32 for discovery and 1.49 for joint follow-up.

**C.3.c. Expected Outcomes of Aim 2**
We anticipate that this study will identify common germline genetic variants important to the risk of SAEs in patients receiving standardized treatment for stage II-III colon cancer. The use of standardized assessments for SAEs across the three RCTs makes it highly likely that these findings will be valid and more broadly generalizable. We expect these analyses will help inform future functional and translational follow-up studies.

**C.3.d. Potential Problems and Alternative Strategies**

**Heterogeneity, survivor-bias, and rare variants.** Some of the potential problems discussed under Aim 1 may also be present in Aim 2, and will be addressed as described in Aim 1. Differences between therapies in the included RCTs. FOLFOX was administered to all patients in the included RCTs, but these studies did evaluate different interventions (i.e., cetuximab, bevacizumab, celecoxib) with distinct mechanisms of action. Additionally, participants in 80702 were randomly assigned to either 3 or 6 months of FOLFOX, whereas participants in NO147 and C-08 received 6 months of FOLFOX therapy. We will account for treatment differences by using a 2 df test in our primary analysis, and through proposed exploratory analyses; however, we will not be able to directly replicate findings in this study if we observe an association specific to one of the adjuvant therapies. In the event of such findings, we will approach other RCTs or studies with similar adjuvant therapies to provide replication. Adverse events due to other treatments. Our evaluation of SAEs will be limited to those reported for participants in these RCTs. We will be unable to evaluate other SAEs that may have a genetic association with other colon cancer treatments, such as radiation and irinotecan. However, these therapies are not as widely used in the adjuvant treatment of colon cancer. Thus, our focus on FOLFOX has the most direct clinical application. Dose limitation/drop off bias. Patients may have required reduced treatment doses or may have been unable to continue treatment because of toxicities. While many of these patients will be captured in SAE reports, it is possible that their early or persistent adverse responses did not lead to the development of an SAE report of grade ≥3. To examine this bias, we will also consider the SNP profile of RCT patients who discontinued treatment or experienced a dose reduction. Rare toxicities. Because of sample size considerations, we are focusing on the most common severe toxicities. It may be that specific variants are relevant to less common serious toxicities, such as cardiovascular thrombosis or infarction. However, we will examine such outcomes as part of our assessment of overall SAEs. If any striking findings emerge, we can pursue additional evaluation of these associations in future studies.

**C.4 Specific Aim 3**

**AIM 3:** Examine the impact of adding information on germline genetic loci to existing prognostic models for stage II-III colon cancer currently based on patient characteristics and clinical factors, such as age, sex, tumor location, tumor stage, and histological grade. We hypothesize that the addition of host genetic variation data will improve the performance of these models in predicting prognosis among patients who receive standard adjuvant chemotherapy (FOLFOX).

**C.4.a. Rationale**
Current staging methods for colon cancer are inadequate in predicting prognosis, particularly for patients with stage II-III colon cancer. Several prognostic decision models have been developed to address this gap in knowledge, using patient and tumor characteristics to improve prediction of patient outcomes. One of the more widely used models is Numeracy, developed by Dr. Sargent (Co-I: Mayo Clinic) and colleagues. Numeracy includes information on age, nodal status, tumor stage, and grade to predict DFS and OS, and has been formalized into a web-based interface for use by both patients and clinicians (http://www.mayoclinic.com/calcs). This model and another widely available, web-based prognostic model, Adjuvant! (http://www.adjuvantageonline.com), were recently validated in the context of
population-based and RCT colon cancer patient populations, demonstrating good reliability for prognosis prediction, particularly for stage III patients. Although Numeracy and Adjuvant! performed similarly in that validation study, Adjuvant! includes information on gender, comorbidity, and number of examined lymph nodes, in addition to factors included in Numeracy, and is based on survival estimates derived from the Surveillance Epidemiology and End Results (SEER) tumor registry rather than RCTs. Both models have been shown to accurately predict patient-level outcomes -66% of the time. Thus, although useful, there remains a need to further improve the predictive power offered by Numeracy and Adjuvant!.

A few prior studies have considered the contributions of common genetic variation in the context of validated prognostic or risk prediction models. For example, the Gail model for incident breast cancer risk was recently expanded to incorporate genetic information on a limited panel of seven SNPs. We also recently participated in a multi-center effort to develop a statistical model for predicting risk of incident CRC which included 10 susceptibility loci as well as information on age, gender, and family history. For both prior studies, incorporating genetic factors into traditional risk prediction models yielded only modest gains in individualized risk prediction. However, the addition of genetic data to risk prediction models does appear to successfully identify high-risk subgroups which are likely to benefit from targeted interventions, such as screening. Further, previous work has shown that combining genetic factors with clinical characteristics has a greater impact on model performance when applied to a population at higher risk for the endpoint of interest; the overall likelihood of disease recurrence is relatively high (-30%) for patients with colon cancer as compared to the likelihood of developing incident colon cancer in the general population (-7%).

Models like Numeracy and Adjuvant! offer clinical utility by providing more accurate, individualized estimates for risk of disease recurrence or death following colon cancer diagnosis. Newly diagnosed cancer patients consistently report a desire to have more specifically tailored estimates about their prognosis. Thus, expanding these prognostic tools to include information beyond existing colon cancer clinical factors might provide important decision aids for colon cancer patients who are considering adjuvant therapy. More broadly, there is a high (and unmet) clinical need to develop methods of risk stratification for outcomes among patients with colon cancer which could be used to inform who, when, and how to treat patients and tailor treatment decisions for maximizing benefit and minimizing toxicity. Given the substantial costs to the United States health care system associated with colon cancer treatment, "stratifying" or "personalizing" cancer therapy according to risk profile has been identified as a high priority. Thus, our planned prognostic model development will be innovative, important, and computationally tractable utilizing the data from this proposal.

Preliminary Studies. Numeracy was developed by Dr. Sargent (Co-I: Mayo Clinic) and colleagues using a pooled dataset of 3,302 patients with stage II-III colon cancer from seven RCTs, all of which had involved random assignment of participants to receive either FU-based chemotherapy or no therapy after resection. Using information on nodal status (0 vs. 1-4 vs. ≥5 affected nodes), depth of invasion (T-stage T1/T2 vs. T3 vs. T4), and grade (low vs. high), Numeracy has been shown to yield a c-statistic of 0.66 for 5-year DFS. Adjuvant! was originally developed in the setting of early stage breast cancer and has been extended to model CRC prognosis; it further incorporates information on sex (male vs. female), comorbidity (perfect health vs. minor problems vs. average for age vs. major problems), and number of examined lymph nodes (0 vs. 1-3 vs. 4-10 vs. >10). Dr. Sargent and colleagues recently validated these prognostic models in independent datasets from a population-based and an RCT setting. In that analysis, Numeracy and Adjuvant! were similar in terms of overall reliability, with accurate predictions of 5-year OS noted 63-64% of the time in the population-based cohort and 69-70% of the time in the RCT; both models, however, tended to overestimate OS and recurrence-free survival in patients with stage II colon cancer who received adjuvant 5-FU chemotherapy. Dr. Sargent's extensive experience working with these prognostic models will be integral to efforts proposed here. In consultation with Dr. Sargent, Dr. Zheng (Co-I: FHCRC) will incorporate information on germline genetic factors into Numeracy and Adjuvant!. Dr. Zheng and colleagues have developed novel statistical methods for developing and validating prognostic models with censored failure time using multiple cohorts.

C.4.b. Methodological Approach
As in Aims 1 and 2, research conducted under Aim 3 will include data from three RCTs of stage II-III colon cancer (N0147, C-08, 80702). We will construct the prognostic models using 6,557 cases and will validate these models in an independent sample of 1,000 cases pooled from these RCTs.

Statistical Analysis. Analyses for Aim 3 will involve phases of model building and model follow-up. In both phases, we will compare the predictive performance of the existing Numeracy and Adjuvant! models to the predictive performance of these models with added information on common genetic variation. In the model building phase we will evaluate these models with and without genetic data in the context of the combined N0147 / C-08 / 80702 stage II-III study.
population from Aim 1 (N=6,557). In the model follow-up phase we will implement models with and without SNPs selected in model building in a sample of 1,000 additional stage II-III cases from N0147 (N=250), C-08 (N=250), and 80702 (N=500) that we will have reserved from Aim 1 to be used as an independent sample for model validation. In both model building and follow-up phases, we will compare predicted survival estimates from our new models with the actual observed estimates, similar to the strategy used by Gill et al. for validating Numeracy and Adjuvant!, with a focus on whether the new models show improvement over predicted survival estimates from models not including genetic factors. We will conduct separate analyses for the prediction of 5-year DFS and 5-year OS. We will account for censoring with survival analysis methods. Model discrimination will be evaluated using the area under the receiver operating characteristic curve (AUC), which can be defined here as the probability that a randomly selected person who experiences an event (i.e., disease recurrence (DFS) or death (DFS / OS)) during the 5-year period is assigned a higher risk than a randomly selected person who remains event-free. A limitation of the AUC as a measure of discriminatory improvement is its insensitivity to the discriminatory gains associated with important covariates. Therefore, we will supplement our evaluations with an additional measure of Integrated Discrimination Improvement (IDI), which is determined based on observed differences in average 5-year risks among those who do versus do not experience an event during the 5-year interval. In addition, we will calculate Net Benefit (NB) based on the decision curve literature to evaluate the cost-effectiveness of using the model. We will estimate the improvement in terms of predictive capacity for new models compared to the original Numeracy or Adjuvant! models by calculating differences in AUC, IDI, and NB evaluated from the new model (with SNP data) versus original models (not including SNP data). A confidence interval for the difference covering 0 would indicate that the improvement is not statistically significant. If our validation shows promise for the new models including genetic variants, further validation in additional large, independent, cohorts of stage II-III colon cancer patients will be warranted.

**Model building.** In the model building phase of Aim 3, we will first evaluate the Numeracy and Adjuvant! models in the absence of genetic data. Evaluation of the prognostic models will be performed separately for DFS and OS. In a multivariate Cox regression model, we will include the following key clinical variables used by Numeracy and/or Adjuvant!, as described above: age, sex, nodal status, T-stage, grade, comorbidity, and number of examined lymph nodes. To determine the incremental gain in predictive power by adding genetic information to these models, we will subsequently include data on all SNPs that attain a significance level of 5x10^-4 in Aim 1 analyses of DFS to regression models of DFS; we will do the same for regression models of OS. This screening step makes the model selection numerically more tractable while including SNPs that are at least weakly associated with the survival outcomes of interest. We set the significance level to be less stringent than the genome-wide significance level of P 5x10^-5 adopted in Aim 1 as we hope to include SNPs that are informative of prognosis when combined with other SNPs and clinical risk factors, even if they are not independently significantly associated with prognosis. Interactions will also be examined by taking the product of genetic variants and clinical risk factors. We will use a stepwise approach to determine which SNPs and interaction terms to include in the final parsimonious prognostic model, while retaining the key clinical factors originally used by Numeracy and Adjuvant! We will pool the included studies and will use a stratified Cox model to allow for studies having different baseline hazards for model outcomes. This approach has been shown to have comparable statistical efficiency as the meta-analysis approach used in Aims 1 and 2 for single SNP identification. The baseline hazards for survival outcomes will be derived by first estimating hazard rates for each trial then combining them by the usual meta-analysis method with the inverse variance as the weight.

**Model follow-up.** In the model follow-up phase, we will use an independent sample of 1,000 participants from N0147, C-08, and 80702. The 1,000 participants will be randomly selected from the three RCTs and will not be used in the model building stage of Aim 3, or in Aim 1. In this phase, we will again evaluate the predictive power of Numeracy and Adjuvant! with and without the data on genetic factors. In this phase, the modified versions of the prognostic models will include the original clinical factors included in Numeracy and Adjuvant!, plus the SNPs and interactions retained in the parsimonious model generated in the model building phase.

**Statistical power.** As seen in Table 4, with a sample size of 1,000 for model follow-up, we will have sufficient power (0.99) to detect an incremental gain of 10% in AUC beyond the previously reported AUC of 0.66 for 5-year DFS seen for Numeracy. In recent validation analyses, the predictive power of Numeracy and Adjuvant! were similar to each other, and were slightly greater for predicting mortality than for disease recurrence. We chose an incremental value of 10% in AUC as this is a likely threshold by which model performance would need to improve in order to offer significant, meaningful improvement in clinical practice.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>AUC Original models</th>
<th>AUC With SNP data added</th>
<th>Power**</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFS</td>
<td>0.66*</td>
<td>0.71</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>0.66*</td>
<td>0.76</td>
<td>0.99</td>
</tr>
<tr>
<td>OS</td>
<td>0.70</td>
<td>0.75</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>0.70</td>
<td>0.80</td>
<td>0.99</td>
</tr>
</tbody>
</table>

** Based on DeLong 1-sided test for paired AUCs and assumed 5-year DFS=65% and 5-year OS=70%
C.4.c. Expected Outcomes of Aim 3
With the completion of Aim 3 we expect to have estimated the incremental gain in predictive power of validated prognostic models for stage II-III colon cancer when adding information on germline genetic loci. Results from these analyses will help address whether specific germline genetic factors should be evaluated more routinely to enhance risk stratification in existing prognostic models used in colon cancer patients.

C.4.d. Potential Problems and Alternative Strategies
Selection of the study population. A potential concern of the approach described here is that participants in the included RCTs may not be representative of the general population of colon cancer patients due to the stringent eligibility criteria of most RCTs. To improve the robustness of projections for individualized risk to the general colon cancer patient population, we can consider the approach that uses external disease mortality rates from a national registry or other large cohort studies with a population similar to those of interest to obtain the baseline risk. Our proposed work in the three RCTs is a useful starting point for developing more accurate statistical models for individual assessment of prognosis. Ultimately the model should also be recalibrated and validated on population-based cohorts.

Statistical power. To maximize the statistical power for selecting novel loci for prognosis prediction, we will preserve most of our samples for model building. The small size of the validation set (N=1000) will result in lower power to detect relatively small gains (e.g., 5% improvement in AUC) in prediction capacity. However, in light of costs associated with genotyping in the clinical setting, gains of greater magnitude (e.g., 10%), such as those we are well powered to observe, are more likely to be clinically relevant. Results from this relatively small validation will also provide a useful basis to consider whether additional resources are required to pursue future large-scale validation. More comprehensive prognostic models. Our evaluation is based upon existing prognostic models, with the addition of information on the subset of SNPs for which there is evidence of an association with survival outcomes. As new information becomes available, these models may be further enhanced to include additional genetic factors, patient factors, or clinicopathological characteristics. Our research will provide initial proof of principle regarding the value of adding germline genetic information to colon cancer prognostic models.

D. TIMELINE / FUTURE PLANS
Our timeline is summarized in Figure 4. Numerous additional activities will be possible to extend and enhance the work in this proposal. Functional follow-up of identified variants will be forthcoming though the Genetic Associations and Mechanisms in Oncology (GAME-ON) consortia of U19 grants, which includes the ColoRectal Transdisciplinary Study (CORECT), for which Dr. Peters (Co-I of this study) is a Co-PI. CORECT has established an infrastructure for mechanistic studies of specific genetic loci with relevance for CRC progression and mortality. We will also build upon emerging studies, such as The Cancer Genome Atlas, to examine somatic variation in the regions of identified variants for DFS, OS, and SAEs in relation to functional pathways and expression profiles. Additionally, we have recently been funded to pilot whole exome sequencing, and are performing exome chip genotyping in a population-based sample of CRC cases. These efforts will improve our understanding of rare variants that could also relate to DFS, OS, and SAEs, and provides a natural extension to the study of rare variants in the RCT setting. Along these lines, we recognize the potential limitations of using relatively highly selected RCT patient populations. In future studies, we will confirm our findings in non-RCT study populations, including the Genetics and Epidemiology of Colorectal Cancer Consortium of observational studies, of which Dr. Peters is the PI, as well as populations that are more diverse in terms of race and ethnicity. Such follow-up studies will allow us to assess of the generalizability of our findings, and may lead to the identification other significant SNPs relevant to our outcomes that could further augment prognostic models. Finally, we have included in this proposal the only completed NCI-sponsored phase III RCTs of non-metastatic colon cancer that have collected specimens for the isolation of germline DNA suitable for genome-wide SNP genotyping. However, we can extend our approach by following up specific variants with more directed genotyping in other RCT populations in which smaller amounts of DNA can be isolated from normal surrounding tissue in formalin-fixed paraffin-embedded tumor blocks. This will permit examination of identified loci within the context of other treatment regimens, including older RCTs in which there were treatment arms that did not receive any treatment. Finally, our comprehensive assessment of germline genetic variation is highly complementary to ongoing funded studies within the included RCTs examining additional prognostic markers, including intratumoral molecular features. These studies will inform future work examining genomic predictors of molecular subtypes of colon cancer, and the combined influence of genomic and somatic genetic variation. Thus, overall this project represents an efficient and cost-effective approach that effectively complements other concurrent efforts and forms a foundation for extending work into other clinical settings. Taken together, this project represents an unprecedented opportunity to develop several new directions the study of colon cancer clinical outcomes.
FIGURE 4. Timeline for study activities (Years 1-4)

<table>
<thead>
<tr>
<th>Aim 1 &amp; 2</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prepare and ship DNA</td>
<td>Genotype at CIDR</td>
<td>QC</td>
<td>Discovery phase genome-wide analysis and SNP selection</td>
<td>Follow up</td>
</tr>
<tr>
<td>Harmonize clinical/outcomes/toxicity data</td>
<td></td>
<td></td>
<td>Exploratory aims</td>
<td></td>
</tr>
</tbody>
</table>

| Aim 3 | | | Model follow-up |
|-------| | | |
| Model development | Model building incorporating SNP data | | Manuscripts |
INCLUSION ENROLLMENT REPORT

Data for this study has already been collected. We will be using data from 7,557 cases of CRC recruited from multiple different studies. We are using existing samples and data and are not imposing any additional inclusion/exclusion criteria. About half of study participants are male (49.7%). The majority of the study subjects are White (87.9%); other racial/ethnic groups represented include Black or African-Americans (7.1%), Asian (4.1%), Hispanic or Latinos (4.8%) and Native Hawaiian or other Pacific Islander (<1%) and American Indian/Alaska Native (<1%).
HUMAN SUBJECTS

Risks to the Subjects

Human Subjects Involvement and Characteristics. This application proposes no new study participant recruitment. No additional contact with participants is required. Cases from these studies have already been recruited (NCCTG N0147, NSABP C-08, CALGB 80702) or are currently being recruited (CALGB 80702) independent of the proposed project. Thus, we will build upon the resources of these existing studies. This will include the use of previously collected data, including available DNA and/or genome-wide SNP data and clinical data and results from previously run molecular tests. Each participating study has already collected information on toxicity and vital status for their subjects and/or are continuing to collect this information prospectively in accordance with the existing consent for the each study. We will perform genotyping on samples from two clinical trial studies (NCCTG N0147 and NSABP C-08), using existing biospecimens. All work will be done in accordance with policies from individual studies' institutional review boards (IRB).

For the proposed project, we will include approximately 7,557 participants from the existing studies with available data (see targeted enrollment table). We are not imposing any additional inclusion/exclusion criteria. Over half of the participants are aged 60 and over. Just under half of study participants are male. As show in the targeted enrollment table, the majority of the study subjects are White; other racial/ethnic groups represented include Black or African-Americans, Asians, Hispanic or Latinos, Native Hawaiians or other Pacific Islanders, and American Indian/Alaska Native.

Sources of materials. We will use existing data including clinical information. Updated vital status and cause of death will be ascertained from the data study coordinating center for each study. Genotyping will be performed using existing biospecimens (buccal cell, blood specimens, or already extracted DNA).

Potential risks. No risks are anticipated for study participants beyond those already identified in the original recruitment and data collection protocols. Questionnaire information and biospecimens used for this proposal have already been collected. No other risks are anticipated for study participants.

Adequacy of Protection against Risks

Recruitment and informed consent. No active recruitment will take place as part of this study; recruitment has already been completed under the existing studies or is currently ongoing independent of this project. This proposal will use data that has already been collected under IRB-approved protocols. All study participants have consented to the data collection and laboratory analyses proposed as part of their participation in the primary studies. We will apply for IRB approval of this study at the Fred Hutchinson Cancer Research.

Procedures for protection against or minimizing risk. All identifying information is kept in restricted-access password protected files. Only authorized persons have access to confidential data. Personal identifiers, such as name and date of birth, will only be used out of necessity to obtain vital status from each country's national death registry. Only de-identified data will be available to the study investigators. Since no analytic file for this study requires identifying information, all files and each specimen will only include a unique ID number assigned. All biological samples will be anonymized, containing only a unique study ID number. When required we will use third party anonymization to ensure that the biological samples cannot be linked to other analytical files. No individual laboratory results from this project will be released to the participant or become part of their clinical medical record without a separate IRB approved process.

Risks and Benefits to Subject

The risks to participants are minimal, as there is no participant contact. In addition, this study requires no further burden on the participants because informed consents, epidemiologic data, medical records, and any other study data has already been obtained. Though there is no potential for a direct benefit to the participant, the benefits to society may be significant. The information from this study may help clinicians to establish a more accurate prognosis and make more informed decisions regarding treatment and follow-up of colon cancer patients.

Importance of Knowledge to be Gained

The proposed study will greatly increase what is currently known about genetic variants and colon cancer outcomes. The knowledge gained in this study may eventually have implications for how clinicians treat and make prognoses for colon cancer patients. In addition, this research has the potential to identify sub-groups with particularly unfavorable prognoses that are not identified through disease staging. These subgroups may be candidates for future clinical trials of colon cancer treatment and other interventions to improve survival. Furthermore, by exploring differential treatment effects, we may be able to identify sub-groups that do not respond well to a particular colon cancer treatment. This information may be used to motivate further investigations of the best way to treat these less responsive sub-groups. Overall, the knowledge gained through the proposed study will be extremely important from both a scientific and clinical standpoint.
INCLUSION OF WOMEN AND MINORITIES

About half of the cases included in this study are female. The racial and ethnic composition of the study reflects the existing distribution in the participating studies. The proposed work has no additional exclusions based on gender, race, or ethnicity.
TARGETED/PLANNED ENROLLMENT TABLE

This report format should NOT be used for data collection from study participants.

Study Title: Molecular Correlates of Outcomes in Clinical Trials of Colon Cancer
Total Planned Enrollment: 7557

<table>
<thead>
<tr>
<th>Ethnic Category</th>
<th>Females</th>
<th>Males</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic or Latino</td>
<td>183</td>
<td>182</td>
<td>365</td>
</tr>
<tr>
<td>Not Hispanic or Latino</td>
<td>3,611</td>
<td>3,581</td>
<td>7,192</td>
</tr>
<tr>
<td><strong>Ethnic Category: Total of All Subjects</strong> *</td>
<td>3,794</td>
<td>3,763</td>
<td>7,557</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Racial Categories</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>American Indian/Alaska Native</td>
<td>15</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>Asian</td>
<td>156</td>
<td>156</td>
<td>312</td>
</tr>
<tr>
<td>Native Hawaiian or Other Pacific Islander</td>
<td>15</td>
<td>14</td>
<td>29</td>
</tr>
<tr>
<td>Black or African American</td>
<td>272</td>
<td>270</td>
<td>542</td>
</tr>
<tr>
<td>White</td>
<td>3,336</td>
<td>3,308</td>
<td>6,644</td>
</tr>
<tr>
<td><strong>Racial Categories: Total of All Subjects</strong> *</td>
<td>3,794</td>
<td>3,763</td>
<td>7,557</td>
</tr>
</tbody>
</table>

* The "Ethnic Category: Total of All Subjects" must be equal to the "Racial Categories: Total of All Subjects."
INCLUSION OF CHILDREN

No subjects less than 18 years of age are included. Colorectal cancer is rare in children.
VERTEBRATE ANIMALS

Not applicable.
SELECT AGENT RESEARCH

Not applicable.
CONSORTIUM/CONTRACTUAL ARRANGEMENTS

Two sub-award budgets for the following institutions/collaborators are included as part of this proposal. Detailed budgets and budget justifications are included for each institution immediately following this section.

**Massachusetts General Hospital (MGH) — Andrew Chan, MD, MPH**
Dr. Andrew Chan, Associate Physician, is a Principal Investigator for this project. This subcontract covers his effort on this project in order to lead and coordinate the current proposal with Dr. Newcomb. Additionally, it supports a post-doctoral fellow to take the lead role on integration of the CALGB 80702 and work collaboratively on analyses.

**Mayo Clinic — Daniel Sargent, PhD**
This subcontract covers funding for Dr. Daniel Sargent, Professor of Biostatistics, Dr. Steve Alberts, Professor of Oncology and co-chair of the NCCTG Gastrointestinal Cancer Committee, Dr. Steve Thibodeau, Professor of Laboratory Medicine and co-director of Molecular Genetics, and Dr. Qian Shi, Assistant Professor of Biostatistics, as scientific collaborators. These Mayo Clinic investigators will provide clinical, scientific, and biostatistical expertise, and will lead all statistical analyses of clinical trial data from NCI-funded cooperative groups. This subcontract also covers a statistical programmer analyst, as well as DNA extraction and preparation for NCCTG study.
RESOURCE SHARING PLAN

A. Data Sharing Plan
We will publish our results in scientific journals in a timely manner and will present study findings at scientific meetings. We will follow all relevant policies and practices of the individual studies (including the requirement that genotyping results be entered into the study-specific databases) and the Internal Review Boards (IRB) of the study centers regarding human subjects protection and IRB privacy. Data on genotypes will be deposited into the Consortium database after quality-control evaluation has been completed, to be shared among investigators. NIH Intellectual Property guidelines will be followed, consistent with study-specific rules and procedures.

We propose to share data and supporting documentation upon request to outside investigators, after review and approval using existing mechanisms within participating studies and groups (NSABP, NCCTG, CALGB). Data for potential collaborators will be compiled as de-identified datasets to safeguard participant confidentiality, and will be produced at the FHCRC or at NCCTG using relevant software. We will make a standardized dataset available to approved investigators, based on a subset of data from the study database; this dataset will be compiled and available after completion of the stated aims of this study. The requesting scientist will work with one of the proposal investigators to determine specific data to be included. The study/group investigators will review the data requests according to their usual protocol. Should any request for data overlap with analyses planned by internal investigators with in the study/groups, data will be made available after a reasonable time has elapsed to allow the original investigators time to publish results. Costs associated with special handling of the data or additional consulting will be passed on to the requesting party.

B. Sharing Model Organisms
Not applicable.

C. Genome-Wide Association Studies (GWAS)
This grant does not cover funding for new genome-wide data; however, new genome-wide data will be generated in the conduct of this project through funding provided by the Center for Inherited Disease Research (CIDR). For data collected during the course of this study through other grants, we will follow the NIH policy for sharing of data obtained in NIH-supported GWAS (1X01HG006662-01) as outlined in (NOT-OD-07-088; http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html).

The proposed work also makes use of genome-wide scan study data, including:
   a) Data already collected through other grants following the NIH policy for sharing of data obtained in NIH supported conducted GWAS as outlined in (NOT-OD-07-088; http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html.)
   b) Data collected through grants funded by agencies other than NIH.

All of these studies are in accordance with NIH policy, and we will make sure our use of the data is in accordance with each study's existing data sharing policy.