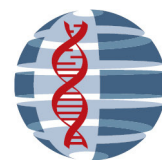


INTERNATIONAL CANCER GENOME CONSORTIUM

GOALS, STRUCTURE,
POLICIES & GUIDELINES

APRIL 2008

FOR UPDATES VISIT:
WWW.ICGC.ORG/POLICIES



ICGC

International Cancer Genome Consortium

Goals, Structure, Policies & Guidelines

TABLE OF CONTENTS

Section	Page Number
A. Introduction	1
B. Consortium Goals	2
C. Background to the Consortium	3
D. Structure of the Consortium	4
ICGC Funding Members	4
ICGC Research Members	5
Governance	6
Coordination	7
International Scientific Steering Committee	
Data Coordination Center	
Quality Assessment Centers	
E. Consortium Policies and Guidelines	8
E. 1. Informed Consent, Access and Ethical Oversight	9
E. 2. Data Release Policies	15
E. 3. Publication Policy	16
E. 4. Intellectual Property Policy	17
E. 5. Tumor Types and Subtypes	17
E. 6. Quality Standards of Samples	20
E. 7. Study Design and Statistical Issues	22
E. 8. Genome Analyses	24
E. 9. Data Management	28
Appendix: Working Groups, Scientific Planning Committee and Interim Executive of the ICGC	32

International Cancer Genome Consortium

A. Introduction

Cancer incidence and deaths are rising worldwide as a result of the growth and aging of the human population. It is estimated that in 2007 over 12 million new cases were diagnosed across the planet and approximately 7.6 million cancer deaths occurred; these numbers will rise to an expected 27 million new cases and 17.5 million cancer deaths in 2050¹ if our ability to prevent, diagnose and treat cancer does not improve. The consequences of cancer for individuals, their families and society are enormous. Although it is difficult to estimate the financial costs, these are also large, through direct costs to health care systems and indirect costs of lost economic output. There are many etiological factors in cancer including infection, exposure to chemicals (e.g., in tobacco smoke), diet, radiation (e.g., in sunlight), and heredity. While several of these factors are preventable, many are not.

All cancers arise due to alterations in DNA. Some cancer-causing mutations may be present in the germline, are therefore heritable and confer an elevated risk of developing cancer. Many, however, occur over the course of a person's lifetime in individual cells of the body and are known as somatic mutations.

Within each cancer genome, a subset of the somatic alterations are "driver" mutations in "cancer genes" which cause the cancer to develop. The search for cancer genes and the "driver" mutations within them has been a central aim of cancer research for 30 years and more than 300 genes have already been identified in which somatic alterations are associated with cancer. The study of these "cancer genes" has generated most of our biological insights into the process of oncogenesis. Cancer genes and the pathways in which they are involved have been used successfully as the targets for the development of new therapeutic agents.

Cancer genomes also carry "passenger" mutations, which do not contribute to the neoplastic phenotype and are not associated with selective growth advantage. They can, however, constitute a molecular record of each cancer's evolutionary past, reflecting past mutagenic exposures and intrinsic defects of DNA repair. As a result, they can inform powerfully on the etiology of individual cancers.

There exist major differences among cancer types both in the cancer genes, i.e., driver mutations, that are operative and in the numbers and types of passenger mutations found. The current evidence indicates that our understanding of patterns of somatic passenger mutation in cancer is at an early stage and that there are many cancer genes still to be identified. Establishing a complete catalogue of somatic genetic changes in individual cancers will therefore reveal the full set of driver mutations and cancer genes that are operative in each type of cancer. It will also reveal the full set of passenger mutations and hence yield insights into underlying mutational processes, including exposures and DNA repair defects.

The underlying biological diversity of human cancer, even those within the same pathological class, and the multiplicity of biological pathways that may be subverted mean that individual cancers may need different treatments depending on the specific genetic abnormalities within them. Recent additions to

¹ Garcia et al, Global Cancer Facts & Figures 2007, Atlanta, GA, American Cancer Society 2007.

the therapeutic arsenal used to treat cancer reflect this trend. These include therapeutic monoclonal antibodies for breast cancer with amplification of *HER2/NEU*, and small molecule inhibitors for chronic myeloid leukaemia carrying a *BCR/ABL* translocation or for lung cancer with *EGFR* mutations. Understanding of and attention to the underlying genetic diversity in cancer is, therefore, likely to increase the success of new cancer modalities in the future.

The sequencing of the human genome allowed the identification of the full set of protein coding and other classes of gene. Although this has catapulted the scientific community into a new era of disease research, the systematic, genome-wide identification of all somatic abnormalities in large numbers of individual cancers has not been technically feasible until recently. It is now possible to contemplate the complete cataloguing of genetic alterations in different types of cancers, with the expectation that our current ability to classify tumors will be refined and improved by classification according to the mutational profiles of each tumor. Combined with the development of rational therapies on the basis of new understanding of the genomics of the individual cancers, many new therapeutic opportunities will become available.

Creating a catalogue of mutations in cancer is an ambitious project. The International Cancer Genome Consortium (ICGC) has been organized as an international effort to harmonize the large number of projects that are now, or shortly will be, underway that have the common aim of elucidating comprehensively the genomic changes present in many forms of cancers that contribute to the burden of disease in people throughout the world. The expectations are that the outcome of the research carried out by the members of the ICGC will be extensive. First and foremost will be compendiums, available to the world-wide research community, of genomic alterations in many cancer subtypes. Second, valuable information on the methods utilized by ICGC members to produce, analyze, and integrate large genomic datasets related to cancer will be made immediately available. Third, it will be possible to compare molecular differences in specific cancer subtypes found in different geographic areas. The ICGC will facilitate communication among the members and provide a forum for coordination, with the objective of maximizing efficiency among the scientists working to understand, treat, and prevent these diseases.

B. Consortium Goals

Primary Goals

1. Coordinate the generation of comprehensive catalogues of genomic abnormalities (somatic mutations) in tumors in 50 different cancer types and/or subtypes which are of clinical and societal importance across the globe.

Ensure high quality by defining the catalogue for each tumor type or subtype to include the full range of somatic mutations including single-nucleotide variants, insertions, deletions, copy number changes, translocations and other chromosomal rearrangements, and to have the following features:

- **Comprehensiveness**, such that most cancer genes with somatic abnormalities occurring at a frequency of greater than 3% are discovered;
- **High resolution**, ideally at sequence-level resolution;

- **High quality**, using common quality standards for pathology and technology;
 - **Based on control data**, generated from matched non-tumor tissue, to distinguish somatic aberrations from inherited sequence variants.
2. Generate complementary catalogues of transcriptomic and epigenomic datasets from the same tumors.
 3. Make the data available to the entire research community as rapidly as possible, and with minimal restrictions, to accelerate research into the causes and control of cancer.

Secondary Goals

4. Coordinate research efforts so that the interests and priorities of individual participants, self-organizing consortia, funding agencies and nations are addressed, including use of the burden of disease as a criterion for target selection, and the minimization of unnecessary redundancy in tumor analysis efforts.
5. Support the dissemination of knowledge and standards related to new technologies, software, and methods to facilitate data integration and sharing with cancer researchers around the globe.

Given the many uncertainties in a project of this scope - such as in predicting the speed at which technologies will evolve, the time that will be needed to acquire sufficient numbers of high-quality samples for some of the proposed tumor types, and other factors - the ultimate timeframe to generate data for 50 tumor types cannot be accurately projected, but it is anticipated to take several years. On the other hand, data from some of the tumor projects already underway will actually begin to be available soon after launching the ICGC.

C. Background to the Consortium

Following the provision of the reference human genome sequence, the potential for using systematic genome wide screens for the understanding of cancer biology prompted a number of groups around the world to embark on efforts at comprehensive genomic characterization of tumors. The more recent advent of new sequencing technologies has made implementation of this approach on a large scale a realistic possibility in the near future. Therefore, cancer and genomic researchers and funding agency representatives from 22 countries gathered in Toronto, Canada in October 2007 to discuss strategies to accelerate the comprehensive study of cancer genomes. Attendees agreed that genomic technologies are approaching the stage at which cancer genome analysis will be feasible on a comprehensive and high-throughput scale. Meeting participants enthusiastically endorsed the launching of an international consortium to globally pursue this goal in a coordinated manner.

In recognition of numerous challenges that are specific to each tumor type (and subtype), it was agreed that the level of organization on which cancer genomics within the ICGC will be approached is at the specific cancer type or subtype.

An Executive Committee (EXEC) for the preparation of the ICGC was established, with representatives of funding agencies from Australia, Canada, China, India, Singapore, the United Kingdom, the United States (the National Cancer Institute and the National Human Genome Research Institute) and the European Commission (Observer Status). Dr. Tom Hudson (President and Scientific Director, Ontario Institute for Cancer Research) agreed to provide Secretariat functions for the EXEC. A Scientific Planning Committee (SPC) composed of leading scientists in the fields of cancer, genomics, ethics, and bioinformatics research was also composed. These committees, which were augmented by focused working groups (Appendix), developed the following proposal to provide funding agencies and the scientific community sufficient information to allow them to determine their interest and ability to participate in the Consortium.

D. Structure of the Consortium

The ICGC is a confederation of members that share the common goals and principles described in this document and have agreed to work in a coordinated and collaborative manner within a defined structure.

Members consist of Funding Members and Research Members, each of which is an individual or allied group that will provide a level of funding or scientific expertise sufficient to undertake a **Cancer Genome Project** involving characterization of a minimum of 500 unique cases for each cancer type or subtype. Each Member will have responsibility for financially or scientifically supporting a minimum of one Cancer Genome Project. Research Members will need to have existing or committed funds from an ICGC Funding Member.

It is recognized that, at the outset, potential Funding Members may not yet have designated funds available to support a Cancer Genome Project and thus may be unable to immediately commit the requisite funds. Funding agencies with a prior record of funding large-scale cancer and/or genome projects will be provided an opportunity to join the ICGC in the absence of a qualifying research project for the first year to allow sufficient time for them to follow their normal policies and procedures to secure funds, to plan initiatives of this magnitude, and to make a firm funding commitment.

Categories of membership are defined as follows.

ICGC Funding Members

- 1) Single funding agency; or
- 2) Alliance of organizations, with a representative from a single organization within the coalition appointed to the EXEC. (See **Structure**, below)

To support the characterization of 500 unique cases of one cancer type or subtype to the degree of comprehensiveness described in this document, ICGC Funding Members will be required to provide the equivalent of a minimum of \$20 million US in total on such a project, distributed over 5 years, for operations (salaries, consumables, etc.), excluding overhead/indirect costs and equipment. It is recognized that some countries may have lower research costs, or may be able to provide material

contributions (such as specimens) that may offset the level of commitment. Guidelines will be developed to evaluate the value of “in-kind” or lower contributions, with a general principle that the responsibility will rest with the funding organization to ensure that the level of support will be sufficient to mount a cancer genome project that will meet the guidelines of the ICGC.

Funding organizations, bodies or groups that want to join the ICGC as a Funding Member, can self-nominate to the ICGC Executive Committee (described below) which has the responsibility for review and approval of nominations. To become an initial ICGC Funding Member (or ICGC Founder), nominations must be received before September 1, 2008. Additional funding agencies are encouraged to become Funding Members in the future, as they become ready to contribute to the ICGC and adopt the Consortium’s policies and guidelines.

ICGC Research Members

To join the ICGC as a Research Member, nominations must originate from an ICGC Funding Member that will provide support to the research organization. Research Members will have the demonstrated capability and capacity to support a Cancer Genome Project and will perform ICGC-affiliated cancer genome research according to the set of commonly agreed-upon policies and guidelines described in this document. Nominations are reviewed and approved by the Executive Committee. Such organizations will need to have existing or committed funds from an ICGC Funding Member.

Research Members can be:

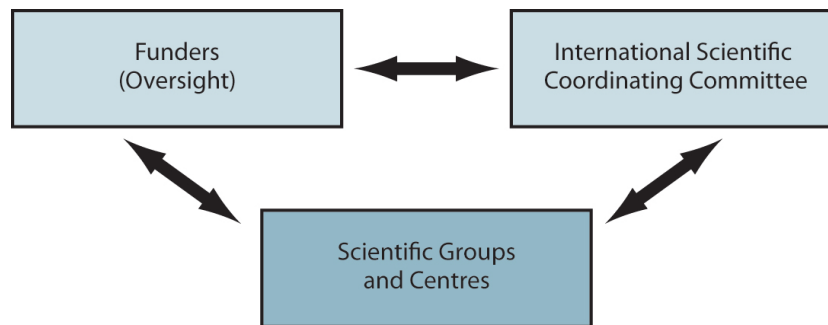
- a) A research center or network of national or international research groups organized to acquire and analyze samples for one or more cancer genome projects;
- b) A genome, cancer, clinical, ethics, bioinformatics (or other) center which contributes significantly to the operations of several cancer genome projects.

Given that these organizations will likely have different structures, and include many investigators, clinicians, scientific managers, as well as clinical and technical staff, each organization will be asked to nominate representatives to participate in ICGC coordination activities, such as the International Scientific Steering Committee, working groups, workshops, and ICGC meetings.

Structure

A distributed model for the organization of the ICGC has been selected as most appropriate for the success of this project. This model has been successfully used in other international genome projects, where high standards and policies have been determined at the outset, and acceptance and adherence were prerequisite for joining. The model, illustrated in Figure 1, relies on the interaction among funders (providing oversight), an international scientific steering committee (setting guidelines) and scientific groups and centers (sample providers and data production centers involved in data production, quality assessment and data management). The strength of the Consortium’s structure rests not only with its component parts but also in the bi-directional flow of information between the groups.

Figure 1: Structure of Consortium



Given the diversity of organizations that will be involved in the ICGC and the fact that most are independently governed, it is understood that in addition to their participation in the Consortium most of the organizations will conduct activities in cancer and genome research that are outside the scope of the ICGC.

Governance

Oversight of the ICGC will be provided by an EXEC, constituted of individuals nominated by ICGC Funding Members. The EXEC will:

- review and accept nominations of new Members;
- work closely with the International Scientific Steering Committee;
- revise or adopt new recommendations related to ICGC policies;
- monitor progress, data quality, and data accessibility across projects;
- periodically report progress to funding agencies;
- provide a forum to discuss potential overlaps that may arise between projects and negotiate solutions;
- provide a forum to resolve issues that may arise;
- decide about recruitment of consultants or establish expert committees on issues related to science, law, intellectual property, ethics, funding, communications, etc.;
- develop a communications strategy, designate communication leader(s), and assure active consultation of all ICGC stakeholders. The importance of ICGC activities will not be overstated, given that the practical benefits to the public will take time to be realized.

The EXEC that was constituted after the October 2007 meeting in Toronto will act as the Interim EXEC of the ICGC, until a permanent team of committed Funding Members is identified.

INTERIM ICGC EXECUTIVE

Warwick Anderson, National Health and Medical Research Council, Australia (Observer Status)

Cindy Bell and Karen Kennedy, Genome Canada, Canada (Observer Status)

Tom Hudson, Ontario Institute for Cancer Research, Canada

Henry Yang, Chinese Cancer Genome Consortium, China

Jacques Remacle, Patrik Kolar and Iiro Eerola, European Commission (Observer Status)
M.K. Bhan and T.S. Rao, Ministry of Science & Technology, Department of Biotechnology, India
Edison Liu, Genome Institute of Singapore, Singapore
Alan Schafer and Michael Stratton, The Wellcome Trust; Wellcome Trust Sanger Institute, United Kingdom
Anna Barker and Daniela Gerhard, National Cancer Institute, United States
Francis Collins, Jane Peterson, Mark Guyer and Brad Ozenberger, National Human Genome Research Institute, United States

Coordination

An **International Scientific Steering Committee (ISSC)** will be constituted with the principal investigators of cancer genome projects in the ICGC, the Data Coordination Center, expert pathologists, oncologists and ethicists, and representatives of funding agencies. This group will interact frequently, through phone conferences, e-mail and regular meetings, to:

- act as a science coordinating body;
- evaluate progress;
- address arising issues of a scientific nature, including those related to samples, consent, ethics, quality standards, evolving technologies;
- exchange protocols, standard operating procedures;
- establish temporary or permanent subcommittees that would be assigned focused tasks;
- establish QC standards.

A **Data Coordination Center (DCC)** will manage data flow from projects and centers to the central ICGC database, public repositories, quality assessment, curation and data releases (see details in section E.9 Data Management). The DCC will provide regular progress reports to the EXEC and ISSC.

Quality Assessment Centers

Quality assessment of the samples used in cancer genome projects is critical to the success of the project. To that end, the Consortium may consider establishing quality assessment centers. The issue of 'round robin' style versus 3rd party quality assessment will require further discussion, as well as mechanisms for funding such activities.

Coordination Support

Staffing will be committed to help manage the operations of the ICGC committees.

E. Consortium Policies and Guidelines

In planning the ICGC, the Scientific Planning Committee recognized the importance of generating a document that would be communicated widely, and contain sufficient information to allow funding agencies and scientists in many countries to make decisions on future participation. Incomplete scientific knowledge (such as tumor heterogeneity for many cancers), rapidly evolving technologies e.g., next generation sequencing technologies, diversity of funding mechanisms, and differences across nations in regards to informed consent and/or sharing of samples across international boundaries are examples of issues that were considered by the committee and its working groups. The approach adopted by the planning teams has been to define a limited number of principles that are central to participation in the project, and provide recommendations to readers based on what is considered “best practices” at the time of writing this document. The authors attempted to discriminate essential from recommended principles using the terms “policy” and “guideline”.

What is a consortium policy?

A consortium policy is a principle which consortium members agree to follow, during the course of the project. Although policies will likely be long-lasting, the ICGC will periodically review its policies.

POLICIES are highlighted in grey.

What is a consortium guideline?

Consortium guidelines refer to recommendations made by ICGC working groups that offer advice as to what is believed to constitute “best practices” at a given time. Given the rapid evolution in technologies or new knowledge gleaned from the data generated by ICGC or other groups, it is expected that guidelines will evolve. It is also expected that approaches will need to vary based on tumor types, local laws, or other factors. In such cases, it is expected that ICGC members will be able to compare and explain differences in approaches, relative to ICGC guidelines. The ICGC has chosen to make most of its recommendations as guidelines rather than policies to allow flexibility in approaches and promote innovation. In this document, guidelines are often written in blue-shaded boxes.

In this first document prepared by the ICGC, the authors strived to differentiate recommendations that are policy from those that are guidelines (even if some issues are clearly a mix of both). It is up to individual projects that join the ICGC to declare a clear plan, e.g., samples, criteria for being a sample, exons used, quality control, etc.

Over time, the ICGC will generate best practices documents that will describe the current state-of-the-art, and propose modifications of the guidelines.

E. 1. Informed Consent, Access and Ethical Oversight

1. Informed Consent

ICGC proposes that certain **Core Bioethical Elements** be respected by all members as a precondition of membership. These elements apply both to the prospective collection of cancer and other samples and to consent surrounding retrospective research using previously stored samples. Following these policies are guidelines that ICGC-member projects should consider in matters related to consent. ICGC-member projects will be responsible for carrying out these policies and guidelines. Nevertheless, ICGC acknowledges that the informed consent process used by ICGC members will necessarily differ according to local, socio-cultural and legal requirements.

POLICY: ICGC membership implies compliance with Core Bioethical Elements for samples used in ICGC Cancer Projects

1.1. Prospective Research

Core Bioethical Elements:

For prospective research, ICGC members should convey to potential participants, *that*:

- The ICGC is a coordinated effort among related scientific research projects being carried on around the world
- Participation in the ICGC and its component projects is voluntary
- Samples and data collected will be used for cancer research, which may include whole genome sequencing
- The patient's care will not be affected by their decision regarding participation
- The samples collected will be in limited quantities; access to them will be tightly controlled and will depend on the policy and practices of the ICGC-member project.
At least a small percentage of the samples may be shared with international laboratories for the purposes of performing quality control studies
- Data derived from the samples collected and data generated by the ICGC members will be made accessible to ICGC members and other international researchers through either an open or a controlled access database under terms and conditions that will maximize participant confidentiality
- Those accessing data and samples will be required to affirm that they will not attempt to re-identify participants
- There is a remote risk of being identified from data available on the databases
- Once data is placed in open databases, that data cannot be withdrawn later
- In controlled access databases the links to (local) data that can identify an individual will be destroyed upon withdrawal. Data previously distributed will continue to be used
- ICGC members agree not to make claims to possible IP derived from primary data
- No profit from eventual commercial products will be returned to subjects donating samples

Box 1. ICGC guidelines for information that should be provided to participants regarding prospective research (ICGC acknowledges that the informed consent process used by ICGC members will necessarily differ according to local, socio-cultural and legal requirements):

- ICGC administration, oversight, funding, duration, ethics and scientific approvals and contact persons;
- Who will be recruited and the approach;
- Procedures involved in participation, including any physical and psychological ‘risks’
- Information on the kinds of samples and data that will be collected;
- Protections in place ‘locally’ to ensure the confidentiality of samples and data;
- Research uses of data (ICGC members are encouraged to seek the broadest level of consent that is appropriate at the local level; e.g., “cancer and related research; cancer and other disease-related research”);
- Whether access to samples will be available for purposes such as validation, quality control, research, etc.;
- Whether access to medical/administrative health records will be sought;
- Whether information regarding participation will be included in medical records;
- Provided it is agreed at recruitment, if clinically important and validated findings emerge during the initial recruitment and screening phase, or in the early research, attempts will be made to pass this information back via the clinician, by whatever mechanism may be agreed at the local level;
- Information on whether or not compensation/reimbursement is available;
- Withdrawal procedures, such as sample retrieval and/or destruction and data coding and anonymization procedures;
- Ownership of samples;
- Prospects for third-party commercialization and intellectual property procedures;
- Purposes for which the uses of data and samples will not be allowed (if required to be named by country);
- How information on the general results of the research will be disseminated;
- Who participants can contact regarding their concerns.

1.2. Retrospective Research

1.2.1. Living Individuals

When conducting research involving samples or data collected previously and the individuals are still living, if existing consents are already in place that will allow ICGC research, the research should proceed without re-contacting those individuals. If appropriate consents are not in place, these individuals should be re-contacted for their consent to participate in the ICGC.

In this situation, the **Core Bioethical Elements** and the guidelines in Box 1 will remain the same as for prospective research, with the exception that where the individual is no longer a patient, there will not be a concern that their care could be affected by participation.

1.2.2. Deceased Individuals and Anonymized Collections

Retrospective research can also be carried out using other sources of samples and data such as: samples and data from the deceased and anonymized collections. Using these sources raises different issues regarding consent; guidelines are provided in Box 2.

The **Core Bioethical Elements** for research involving samples and data from deceased individuals are *that*:

- Where required by law or ethics, consent should always be obtained from the families of a deceased individual if their samples and data are to be used; if re-consent is not required, however, ethics review is sufficient
- Ethics committee review should be sought for all research proposing the use of existing sample and data collections
- Existing collections are a limited and valuable resource; access to them will be tightly controlled.

Anonymized samples cannot at this time be used to retrace individuals or their families.

The **Core Bioethical Element** is that such samples can be used, subject to the removal of any identification or possible combination of factors allowing re-identification and ethics review.

Box 2. ICGC guidelines for retrospective research using existing collections of samples and data (ICGC acknowledges that the informed consent process used by ICGC members will necessarily differ according to local, socio-cultural and legal requirements)

For retrospective research using identifiable collections from the living:

- If existing consent allows for ICGC research, samples and data may be used without further consent;
- If consents are not in place, re-consenting should be pursued (see Box 1);
- If seeking consent is judged by the researchers to be impractical or likely to cause distress to individuals, ethics committee approval may be sought to waive consent requirements.

Criteria for a waiver may include that the research:

- Poses minimal risk to the individuals;
- Does not violate individuals' rights;
- Has privacy and confidentiality protections in place;
- Is important and cannot be conducted in any other manner.

For research using identifiable collections from the deceased:

- The wishes of the deceased regarding the use of their coded samples and data should be taken into consideration;
- Relatives of the deceased, if known, may be consulted regarding the wishes of the deceased;
- Research may proceed without consent if a waiver is provided by an ethics committee or if permitted by national legislation and policies;
- If samples and data are anonymized, research may proceed without consent.

1.3. Access

In addition, the nature of the data produced by ICGC members, including prospective cohorts of cancer patients, substantial clinical annotation and extensive genomic data, raises important human subject privacy protection issues. The patient/individual protection policies developed for ICGC are designed to balance two important goals: to facilitate investigations of genomic changes related to cancer and, at the same time, to respect and protect the patients/individuals whose data and materials have been or will contribute to ICGC-member projects. It is technically possible that genomic information (DNA sequence, genotype) generated by the projects comprising the ICGC could lead to identification of an individual if similar specimen data from that person (or a blood relative) were obtained from a third-party database and correlated. There is also a risk of individual identification by computer-based analysis of the clinical data in conjunction with, for example, third-party demographic and healthcare management databases. This potential identification could then publicly link the individual to his/her clinical information collected by the participating projects, and could lead to social risks such as discrimination or loss of privacy.

POLICY: To minimize the risk of patient/individual identification, the ICGC has established the policy that datasets be organized into two categories, open and controlled-access.

Table 1 includes a list of data elements and the data access category within which they will be available.

The first category, Open Access Datasets, will be publicly accessible and contain only data that cannot be aggregated to generate a dataset unique to an individual. The second category, Controlled Access Datasets, will contain composite genomic and clinical data that are associated to a unique, but not directly identified, person.

ICGC Open Access Datasets	ICGC Controlled Access Datasets
<ul style="list-style-type: none"> • Cancer Pathology <ul style="list-style-type: none"> ○ Histologic type or subtype ○ Histologic nuclear grade • Patient/Person <ul style="list-style-type: none"> ○ Gender ○ Age range • Gene Expression (normalized) • DNA methylation • Genotype frequencies • Computed Copy Number and Loss of Heterozygosity • Newly discovered somatic variants 	<ul style="list-style-type: none"> • Detailed Phenotype and Outcome Data <ul style="list-style-type: none"> ○ Patient demography ○ Risk factors ○ Examination ○ Surgery ○ Drugs ○ Radiation ○ Sample ○ Slide ○ Specific histological features ○ Protocol ○ Analyte ○ Aliquot • Gene Expression (probe-level data) • Raw genotype calls • Gene-sample identifier links • Genome sequence files

Table 1. Listing of data categories and level of access restriction on those data.

An International Data Access Committee (IDAC) will be established as a policy-making group. It will develop the additional policies by which investigators can obtain access to controlled data, provide oversight to any ICGC member projects that will have the responsibility to review requests for such data, and monitor compliance by bodies that are authorized to distribute ICGC data, and users of the controlled data. The IDAC will have broad geographic representation (of whom 50% will be non-ICGC members) and will include individuals representing the ICGC Executive, experts in ethics, databases and international law, cancer survivors, potential users of the data, and other independent lay persons. The IDAC will ideally have fewer than 20 members.

As proposed in Box 8 of this document (Additional guidelines for ICGC data management and security), authorizations to access controlled data will be broad, so that authenticated users will get permission to obtain access to controlled data generated from all samples studied by any participating cancer genome project (as the feasibility of providing permissions to datasets originating from single or partial subsets of participating centers has been determined to be unworkable in the context of the ICGC).

The IDAC will also develop guidelines for practical approaches to providing qualified investigators with access to controlled data. In doing so, it will consider mechanisms and tools that have been already in use by other organizations that distribute controlled datasets to international scientists (for example, the U.S. TCGA project or the Wellcome Trust Case Control Consortium). Potential users and their institutions will be required to submit Assurance Agreement forms that include:

- a written description of the purpose of the research to be done;
- an agreement to not to try to identify or contact the donor subjects;
- agreements not to redistribute controlled access data; and
- plans to destroy controlled access datasets once they are no longer being used.

Interested users and institutional officials who are authorized to make legally binding agreements for the institution will have to provide a signed statement agreeing to adhere to the conditions that will be recommended by the IDAC. Investigators will need to agree to regular review and renewal requested by the IDAC for such authorization and in cases when they move to new institutions. No authorization will be granted to scientists or other individuals that are not supported by an institution that assumes responsibility to fulfill the terms of the Assurance Agreement.

The implementation of a data access process will be the responsibility of ICGC Funding Members, i.e., the ICGC Executive Committee, who will establish policies and processes based on the policies and guidelines developed by the IDAC for the Consortium. The management, i.e., receipt, review and approval, of requests for data produced by ICGC members, whether directly or through third-party data repositories wishing to redistribute cancer genome Controlled Access Data ideally will be determined after further consultations. It is anticipated that permissions will be obtained across all jurisdictions to allow establishment of a centralized authentication mechanism at ICGC franchise databases. Delegated organizations will maintain records of all requests for controlled data, authentication procedures, authorizations or denials, and report these semi-annually to the IDAC. Unusual requests or unexpected issues will be referred to the IDAC as they arise.

E. 2. ICGC Data Release Policies

A guiding principle of research funded by public (governmental and charitable) organizations alike is to maximize benefit to the public while, at the same time, protecting the interests of sample donors and their relatives. The data release policies of the ICGC, and of a number of genomic projects before it, are intended to achieve these dual objectives.

Responsibilities of data producers and users

POLICY: The members of the International Cancer Genomics Consortium (ICGC) are committed to the principle of rapid data release to the scientific community.

This principle was first implemented during the Human Genome Project and has been recognized as an extremely beneficial innovation of that and subsequent large-scale genomic analyses, because it accelerates the rapid translation of output data to scientific knowledge. At a meeting in Ft. Lauderdale, FL, which was co-sponsored by the Wellcome Trust and the US National Human Genome Research Institute, in January 2003 (see meeting report at http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf), the concept of rapid data release by genomic sequence data producers was reaffirmed and the attendees strongly recommended adoption of the practice for other types of data produced by **community resource projects** (defined as projects initiated by their developers and funding agencies for the primary purpose of generating a public resource that will be used by a broad range of investigators and others to further the understanding and conquest of disease). The attendees recognized that sustaining the practice of rapid, prepublication data release requires that the interests of all involved - the data producers, data users, and funding agencies - be addressed (in the case of disease-oriented research, patients and advocacy groups are additional stakeholders). They emphasized the need for responsible behavior on the part of all parties and for incentives to induce all concerned to act in ways that will lead to the most social benefit. Data producers are recognized to have a responsibility to release data rapidly and to publish initial global analyses in a timely manner. Of equal importance is the responsible use of the data by end-users, which is defined as allowing the data producers the opportunity to publish the initial global analyses of the data, as specifically articulated at the outset of the project, within a reasonable period of time.

The members of the ICGC agree to identify the projects they support and carry out for the comprehensive genomic characterization of human cancers as a set of community resource projects. Data producers, by explicit agreement as members of the ICGC, acknowledge their responsibilities to release data rapidly and to publish initial global analyses in a timely manner. Similarly, funding agencies acknowledge their role in encouraging and facilitating rapid data release from cancer genome projects.

Data Standards for Data Releases

The ICGC will establish a well-articulated description of a first-level verification standard for each data type produced by Consortium members. ICGC members will release, to an appropriate public database, data obtained in experiments at the time that this standard is met. In most cases, it is anticipated that additional efforts for further verification and validation of the data will be carried out, but these will not delay the initial release of data. The ICGC acknowledges that releasing preliminary data may not be the

first choice of the data producers. However, the ICGC members understand that such data can be useful to the broader scientific community and, ultimately, to cancer patients, so that this policy best serves the objective to enable all qualified investigators to apply the collective intellect to the study and control of cancer. All data will be accompanied by prominent caveats to notify users of the level of verification of the data and that frequent data release and updates will be forthcoming as further validation and analyses are performed.

E. 3. ICGC Publication Policy

POLICY: The individual research groups in the ICGC are free to publish the results of their own efforts in independent publications at any time (subject, of course, to any policies of any collaborations in which they may be participating).

In their individual papers, Consortium participants will not be restricted to describing the methods developed for the project, but can and should expand into describing biological insights that arise from their analyses. To facilitate comparison of data among different groups participating in the ICGC, all publications by Consortium members should include explicit data on quality metrics, possibly including a common reference set of analytes agreed upon by the Consortium, e.g., nucleic acids derived from a common cell line or other source; such papers should also explicitly include a statement that the quality metrics are those that have been adopted by the ICGC (to promote their wide acceptance across the broad research community).

Users of Consortium data, whether members of the Consortium or not, should be aware of the publication status of the data they use and treat them accordingly. For example, all investigators, including other Consortium members, should obtain the consent of the data producers (this term includes clinical contributors and other members of an ICGC collaborative team) before using unpublished data in their individual publications, and the data producers should not unreasonably withhold this consent.

ICGC members will not have privileged access to data from other members of the Consortium. Rather, all data shared by the Consortium members will be obtained from the data that has been released to public databases.

Investigators outside of the ICGC are free to use data generated by ICGC members, either en masse or specific subsets, but are asked to follow the guidelines developed at the Ft. Lauderdale meeting. Specifically, data users should cite the source of the data and should acknowledge the clinical contributors and the data producers from the ICGC. In addition, the data users are asked to recognize the interests of the data producers to publish reports on the generation and analysis of their data, as described previously. Datasets from ICGC members are released to public databases as pre-publication data and remain unpublished until they appear in peer-reviewed publications. Outside investigators who perform an in-depth analysis of data from ICGC members and are interested in publishing a report before the data producers do so should discuss their results with the data producer(s) and are encouraged to establish collaborations. However, ICGC members are not required to collaborate with any outside investigators. All investigators, through their roles as journal and grant reviewers, should enforce a high standard of respect for the scientific contribution of the data producers.

This description of the ICGC data release policy is directed primarily at issues concerning the use of Consortium data in scientific publications. The intent of the policy is to accelerate the use of the data by the global scientific community while, at the same time, allowing the data producers to get appropriate scientific credit for their work through publications. To facilitate this goal, the data producers agree not to restrict the use of the data by others, while the data users are encouraged to act in a manner that is consistent with this policy providing unrestricted access to pre-publication data.

E. 4. ICGC Intellectual Property Policy

The objective of ICGC policy regarding intellectual property (IP) policy is to maximize public benefit from data produced by the Consortium. It is the view of the ICGC members that this goal is achieved if the data remain publicly accessible without any restrictions.

POLICY: All ICGC members agree not to make claims to possible IP derived from primary data (including somatic mutations) and to not pursue IP protections that would prevent or block access to or use of any element of ICGC data or conclusions drawn directly from those data.

Users of the data (including Consortium members) may elect to perform further research that would add intellectual and resource capital to ICGC data and elect to exercise their IP rights on these downstream discoveries. However, if patents are pursued on such “downstream” inventions, ICGC participants and other data users are expected to implement licensing policies that do not obstruct further research; see for example the U.S. National Institutes of Health’s document on “Best Practices for the Licensing of Genomic Inventions” (http://www.ott.nih.gov/policy/genomic_invention.html).

E. 5. Tumor Types and Subtypes that will be studied by the ICGC

The aim of the ICGC is to provide a comprehensive description of the somatic genomic abnormalities present in the broad range of human tumors. Given our current knowledge of the heterogeneity of tumor types and subtypes, the ICGC set a goal of coordinating approximately 50 projects, each of which will generate the genomic analyses outlined elsewhere in this document on approximately 500 cancer samples of each class. It is well recognized, however, that cancer is highly heterogeneous and hundreds of types/subtypes can be defined. Therefore, the stated goal of 50 ICGC projects is not intended to, and cannot, exhaustively cover the full spectrum of cancer types.

With the overall principles of the ICGC in mind, the Clinical and Pathologies Working Group discussed extensively the issues pertaining to selection of specific tumor types/subtypes and defining the criteria/parameters for each project. Recognizing the varying impacts of different cancers in different countries or global regions, the working group recommended that the ultimate justification for an ICGC Project on a specific cancer type will rest with the collaborative groups of pathologists, genomicists, clinical oncologists, cancer geneticists, cancer biologists and epidemiologists that will be proposing each project.

In selecting each project, ICGC members should articulate its importance and relevance based on the public health impact of a cancer type and unmet clinical need. The potential for novel insights that might broadly inform cancer research should also be considered. For example, testis cancer is unique amongst

solid tumors in its high cure rate by conventional chemotherapy when widely metastatic. Depending on the rationale, the subtype or subtypes being proposed may be defined on pathological, molecular, etiological or geographical differences. The scientific merit of each definition of a tumor type for ICGC must be sound and technical feasibility should be addressed. In the case of tumor classes with multiple subtypes characterised by distinct clinical / biological behaviours, a balance may be drawn between studying 500 cases of exclusively one subtype and hence providing optimal power for that subtype, versus including multiple subtypes with lesser power for any one but enabling informative comparisons.

The ICGC aims to study cancers of all major organ systems including central nervous system, hemopoietic and lymphoid tissue, head, neck and nasopharynx, skin, lung, breast, esophagus, stomach, colon, rectum, kidney, bladder and urinary tract, soft tissues, bone, pancreas, gall bladder and biliary system, liver, ovary, uterus, cervix, endocrine tissues, testis and prostate. It is also envisaged that the studies will cover adult and childhood / adolescent cancers, for example neuroblastoma, pilocytic astrocytoma, medulloblastoma, osteosarcoma and childhood leukemias.

Box 3. Guidelines for the selection of cancer genome projects

The Clinical and Pathologies Issues Working Group used the following examples to illustrate some of the issues which will affect the selection of cancer genome projects:

- For many common cancer classes, there are well recognized, clearly defined histopathological or molecular subtypes with differing etiologies or geographical prevalence that merit separate projects, for example transitional cell and squamous carcinomas of the urinary bladder or squamous and adenocarcinomas of the esophagus. Indeed, the classical histological subtypes of some very common cancers, for example adenocarcinoma, squamous carcinoma, small cell carcinoma and large cell carcinoma of the lung are already known to harbour genetic differences. Therefore, each of these might reasonably be represented by a separate ICGC project. For other common cancers, subclassification has recently improved and requires characterization beyond conventional histopathology. For example, in breast cancer the classification of luminal A ER positive PR positive, Luminal B, triple negative, and HER2 positive cancers based on expression of molecular markers is currently believed to optimally reflect the diverse biology of this disease. Similarly in colon and rectal cancer, the presence or absence of mismatch repair defects and the high or low prevalence of methylation changes may represent informative modes of choosing subclasses for study.

Box 3. Guidelines for the selection of cancer genome projects, continued

- For some cancer classes in which one subtype has a significantly higher incidence or mortality than others, it may be pragmatic to focus exclusively on this subtype, for example clear cell renal carcinoma, papillary thyroid cancer and ductal carcinoma of the pancreas. Alternatively, it may be appropriate to formulate two projects, one based on the most common subclass, for example in ovarian cancer a project on serous carcinoma and a second based on one or more of the rarer subclasses of mucinous, clear cell and endometrioid. Similarly, it may be appropriate to choose a small number of the commoner, better-defined classes of soft tissue sarcoma.
- Among some cancer classes, pragmatic issues of tissue collections may argue for separating different stages of progression into separate projects. For example, patients with androgen independent (AI) or metastatic prostate cancers do not routinely undergo surgery. Therefore collection of AI or metastatic prostate specimens requires a different protocol from collection of primary prostate tumors. Similarly, in malignant melanoma, samples from distant metastases may be easier to obtain than from primary tumors, where collection is constrained by medico-legal issues. Under these scenarios, it is reasonable to propose independent projects on primary and metastatic stages of these cancer types. For other tumor types, it may be important to distinguish subtypes based on their patterns of progression, for example secondary glioblastomas, which evolve from lower grade gliomas, and primary glioblastomas, which arise *de novo*.
- It is recognized that multiple etiologies can underlie cancer development in a particular organ system. Even with similar histopathology, the genomics of each subtype may be different. For example, many hepatocellular carcinomas are associated with viral hepatitis, while others are associated with alcohol-related cirrhosis. It would be reasonable to consider these as distinct entities for ICGC projects. Furthermore, selection and comparison of cancer cases with different prevalences in different geographical regions may be important as these may also reflect distinct underlying etiologies. Examples of this form of classification might include cancers of the oral cavity, gallbladder and biliary tract from high-risk areas and gastric, colon, rectal, lung and nasopharyngeal cancers from Asia.
- Tumors of hematological and lymphoid tissues present particular issues. A large number of well established molecular and/or morphological subtypes have been defined, and it is not feasible to conduct full-scale projects on each one. In this context it may be appropriate to select tumor types based on their clinical impact and a relative paucity of information concerning their genetic basis. Examples of such selections might include follicular lymphoma and diffuse large-cell B cell lymphoma, myeloma, myeloproliferative disorders (which represent early stages of neoplastic change), chronic lymphocytic leukaemia, acute myeloid leukaemia with normal cytogenetics, T-ALL and B-ALL.

E. 6. Quality Standards of Samples

Generating collections of high quality tumor samples is likely to be a major challenge of the ICGC. Committed partners and funding agencies will need to invest substantial effort and funds. General guidelines were developed by the Quality Standards of Samples Working Group that could apply to a broad spectrum of cancer subtypes displaying a wide variety of histopathological and clinical characteristics. However, optimal standards may differ considerably between the tumor entities.

POLICY: Every project will adhere to the following four recommendations:

1. A committee of clinical and pathology experts (with representation from different institutions) will be needed to draft and oversee the specific guidelines that will apply for every tumor type or sub-type. These guidelines will have to be made available to all members of the Consortium, and users of the data.
2. Tumor types should be defined using the existing international standards of the WHO (including ICD-10 and ICD-O). If novel molecular subtypes are studied, these should be defined with sufficient detail.
3. All samples will have to be reviewed by two or more reference pathologists. This assessment will need to be performed on stained sections of the very same tissue piece from which biomolecules will be purified.
4. Patient-matched control samples, representative for the germline genome, are mandatory to discern “somatic” from “inherited” mutations. For solid tumors, the mononuclear cell fraction from peripheral blood is the ideal source, while for hematological malignancies skin biopsies or (lymphocytes from patients in remission) are recommended.

Box 4. Guidelines regarding the quality standards of samples

- In the initial projects of the ICGC, it is recommended to begin with tumor entities (and samples) that are the most “homogeneous”;
- In the initial projects of the ICGC, tumor samples should be untreated malignancies, i.e., in general primary and not relapsed, preferably from a single anatomical site of origin and representing a single histological type/subtype (and if feasible, a single histopathologic grade);
- Tumor sample inclusion criteria should require at least 80% of each sample to be composed of viable-appearing tumor cells on histological assessment and less than 20% necrotic cells or normal cells such as inflammatory, immune, or stromal cells, or even pre-malignant (dysplastic) cells. In some tumor types, it might be necessary to adopt a less stringent criterion by the expert panel, but if the tumor cell content is not substantially higher than 60%, physical enrichment procedures need to be considered;
- Histological examination will have to be documented and respective optical images have to be stored and made available to those studying the given tumor entity. Specifically the degree of 1) necrosis; 2) debris; 3) inflammatory tissue; and 4) fibrosis are to be assessed;
- Standard Operating Procedures (SOPs) for freezing samples will be those established by WHO/IARC (2007);
- As a basis for the exchange of tissue specimen between countries with different national regulations that need to be respected, a coordinating rule has been formulated on the basis of the ‘home-country principle’;
- Ideally, 200 mg of solid tissue or 10 million flow-sorted cells (i.e., blood tumors) will be available for each sample. If microdissection is necessary, the number of required cells is still unknown. This aspect needs to be revisited at a later time point;
- Although many types of macromolecules should be isolated, priority should be given to the isolation of high quality DNA (which is also valid for some epigenomic analyses). Isolation of high quality RNA is also recommended;
- The quality of the isolated classes of macromolecules needs to be controlled by standardized procedures used by all members of the ICGC. The choice of these tests will be defined by an ICGC working group;
- Controls for transcriptomic and epigenomic analyses may require site-matched tissue control samples. This aspect must be dealt with in the recommendations of the tumor-specific expert panel;
- The minimum set of clinical variables that must be collected for each tumor sample are:
 - General data (DOB, age, gender, date of diagnosis, presence of metastases, etc.);
 - Diagnostic data (biochemical, cytogenetic, immunophenotypic and other data);
- Acquisition of follow-up information is highly recommended for subsequent interpretation of ICGC data and clinical correlations:
 - Therapies after removal of malignant cells;
 - Response to therapy (EFS, CR, OS, definition of end points of trials).

E. 7. Study Design and Statistical Issues

POLICY: Every cancer genome project should state a clear rationale for its choice of sample size, in terms of the desired sensitivity to detect mutations. The target number of 500 samples per tumor type/subtype is set as a minimum, pending further information to be provided by ICGC members proposing to tackle specific cancer types/subtypes.

The following considerations should be taken into account in planning:

- As a rule of thumb, it is suggested that a tumor/normal collection should be large enough to reliably detect genes that are somatically mutated in 3% of tumors of a given subtype. This is based on the recognition that cancer types can be heterogeneous, with important genes already being found as mutated in 5-10% of samples;
- Based on mathematical analysis, a collection of ~500 samples is needed to reliably detect genes that are somatically mutated in 3% of samples;
- It may not be necessary to fully analyze all genes in 500 samples. Instead, one might use a two-tiered strategy in which (i) genes are studied in a discovery set (N samples) and (ii) a subset of genes that show sufficient frequency of mutations are studied in a validation set (M samples). With N= 100 and M= 400, one still has good power to detect genes that are mutated in 3% of samples;
- While we suggest a detection level of 3% as a rule of thumb for a 'typical' cancer, the detection level should ideally reflect the actual heterogeneity of the cancer subtype. A gene could be mutated in a significant proportion of a subtype, but the overall mutation rate might fall below 3%. In practice, the degree of heterogeneity of a given tumor type is difficult to know in advance.

Nonetheless, some tumor types are known or thought to have more heterogeneous etiologies (for example, sarcomas), which may entail significantly more heterogeneous patterns of genomic and (epi)genetic alterations. In such cases, it could make sense to collect considerably more than 500 tumors.

In other cases, it may make sense to divide cancer types into distinct subtypes based on etiology or biology and, if feasible, assemble collections of each subtype. For example, investigators might be interested in identifying cancer genes associated with distinct subtypes. Examples might include studying smoking-related versus non-smoking-related lung cancers; or hepatocellular carcinomas arising in the setting of alcoholic cirrhosis versus viral hepatitis (B and C) versus helminthic infections versus aflatoxin.

Ultimately, the decision about sample collections must reflect the investigators' best guesses about the underlying heterogeneity and the practical realities of sample collection. It is good to have larger collections at hand, even if only a subset is initially analyzed. The ultimate answer about the degree of heterogeneity will likely come from the genomic data themselves.

Box 5. Mathematical analysis

We briefly outline the mathematical analysis that supports the statements above.

Sample size. To identify cancer-related genes (drivers vs. passengers), one needs to detect genes that are mutated at a higher frequency than the background mutation rate. One has to calculate the probability of observing a given number of somatic mutations in the coding region of (i) a passenger gene in which somatic mutations occur at the background rate and (ii) a driver gene in which somatic mutations occur in 3% of samples.

Background mutation rates can vary between tumors and tumor types, but a typical rate is around 1.5×10^{-6} non-synonymous mutations/base. If we make the simplifying assumption that all genes have a coding region of 1500 bases, this translates to a background rate of 2.25×10^{-3} somatic mutations per gene - or an expectation of ~ 0.625 somatic mutations across a collection of 500 samples). Because there are 20,000 protein-coding genes, some genes will substantially exceed the expectation by random chance. Indeed, one expects that by chance there will be ~ 3.4 passenger genes with ≥ 7 non-synonymous mutations. One must take into account this issue of multiple hypothesis testing – for example, by using a Bonferroni correction.

By contrast, a driver gene in which somatic mutations occur in 3% of samples would be expected to have ~ 15 occurrences among a collection of 500 samples.

If one sets a threshold of 9 somatic mutations across 500 samples to declare significance, the probability that some passenger gene in the genome will exceed this threshold is $\sim 6\%$. By contrast, the probability that a driver gene (3% somatic mutation rate) will exceed the threshold is 98%. If we allow for a missing data rate of $\sim 24\%$ due to incomplete coverage and sensitivity, the probability is 88%.

In summary, a sample of 500 tumors thus provides 88% power to detect a gene mutated in 3% of samples, with a 10% chance of a passenger gene achieving the threshold.

We note that this analysis is only approximate. (i) For example, the genes are assumed to have equal size. More sophisticated statistical models should be used in analyzing actual data from cancer genome projects. (ii) The model uses an average mutation rate per base; it does not reflect differential mutation rates in different nucleotide contexts.

In addition, the sample size analysis focuses only on detection of cancer-related mutations. Different sample sizes may be required, for example, to make accurate risk estimates.

Two-stage design. Using the background mutation rate above, about 4,000 out of the 20,000 genes will have at least one mutation in the first 100 discovery set. Sequencing these 4,000 genes in the remaining 400 samples and requiring a total of at least 9 mutations (in combined discovery and validation sets) only slightly decreases the power to detect a gene which is mutated in 3% of samples, from 88% to 82%. However, this two-tiered strategy can reduce the sequencing costs to 28% of the single-tiered approach.

E. 8. Genome Analyses

The primary goal of the ICGC is to generate catalogues of somatic genomic abnormalities (mutations) in different tumor types and/or subtypes which are of clinical and societal importance across the globe. Assembling a restricted, core set of genomic analyses at the outset is challenging, given that technologies are rapidly evolving and more platforms are in development. It will therefore be necessary to review the recommendations on a regular and frequent basis. It is also preferable to avoid being unduly prescriptive about study designs and platform choices, as it is conceivable that several different technologies and designs could be used to achieve the same ultimate goals.

Ultimately, however, it is critical to the overall success of the ICGC that the datasets obtained from one class of cancer (generated in a particular way) will be directly comparable to the datasets obtained from another class of cancer (even if generated using a different approach/technology). It is particularly important, therefore, that members adhere to quality standards set by the ICGC, including sufficient depth and coverage to detect a high proportion of somatic mutations in each sample that will be interrogated. We outline below the yields of somatic alteration from each cancer expected by the ICGC and a process for evaluating the quality of data generated by different centers.

The following classes of genome analyses are recommended for ICGC membership:

Class 1: Catalogue of Somatic Mutations

POLICY: Genomic DNA analyses of tumors (and matching control DNA) are core elements of the project, and are therefore referred to as mandatory studies.

Ultimately, catalogues for each tumor type or subtype will include the full range of somatic mutations including single base substitutions, insertions, deletions, copy number changes, translocations and other chromosomal rearrangements. It is anticipated that cancer genome studies will expand to include high coverage whole genome shotguns of cancer and normal genomes. **This is an anticipated goal that should be envisaged by all participants joining the ICGC.** A whole genome shotgun design will ultimately provide the optimal, pragmatic strategy and primary output of the ICGC. New generation sequencing technologies are becoming available, the parameters of which make the whole genome shotgun approach potentially a realistic option for application to substantial numbers of cancers within the next 2-5 years.

All sequencing platforms have an error rate. The error rate may in part be mitigated by high genome coverage provided by the whole genome shotgun. However it is likely that errors will remain. We propose that at least 95% of somatic variants listed in the catalogue for each sample should be real. To generate a high quality catalogue of variants may, therefore, require confirmation of somatic variants by a targeted technology, both to exclude sequence artifacts and to eliminate residual private polymorphisms. This curation of each cancer genome may constitute a substantial additional workload to the analysis of cancer genomes and would essentially constitute a “finishing” phase to the generation of the catalogue of somatic variants.

Box 6. Whole genome shotgun analyses (anticipated policy)

The aim of the whole genome shotgun will be to harvest as high a proportion as practically feasible of the somatic mutations present in each individual cancer sample. We propose that at least 80% of the somatic alterations should be identified in each sample and that coverage calculations on each sample should be based on this expectation.

Several issues will need to be understood, however, before specific recommendations on whole genome shotgun designs can be made. In order to provide an almost complete catalogue of somatic mutations in each cancer sample the most important consideration will be the overall depth of sequence coverage obtained, which will determine the sensitivity and specificity of detection of somatic mutations. The coverage required will be influenced by a number of factors including the presence of aneuploidy in the cancer, tissue heterogeneity (normal contamination and tumor subclones), the prevalence of somatic mutations in the cancer, sequence error rates, other data features of the sequencing technology adopted and the proportion of known SNPs.

In order to ascertain which variants are somatic it will be necessary to evaluate variants found in the cancer sample in normal DNA from the same individual. It is likely that this will primarily be obtained through a whole genome shotgun of the normal DNA sample. However, other approaches are not excluded.

Preliminary estimates indicate that approximately 30-fold genome coverage, and possibly more, of the cancer sample will be required. Formal power calculations will be provided to support these predictions and will be adapted in future on the basis of specific information on each platform used.

Whole genome shotgun analyses of cancer genomes may not be feasible for two years or more. Box 7 lists the recommended initial interim goals that are proposed until whole genome shotgun approaches are feasible on thousands of samples. It is expected that members joining the ICGC plan to go beyond the interim analyses, and launch whole genome sequencing when the technologies are shown to be robust and affordable. The ICGC will monitor progress and re-evaluate these guidelines periodically.

Box 7. Interim, large-scale, catalogues of somatic mutations

1. Sequencing of all coding exons and other genomic regions of particular biological interest for point mutations. There are several technologies now available to achieve this goal including enrichment by array pull down or PCR followed by sequencing on one of the new technology platforms. The aim of these analyses would be to find at least 80% of somatic alterations in these regions in each cancer sample. Sequence coverage should be estimated on this basis. The primary targets would be all coding exons / splice sites and microRNAs, followed by regulatory and conserved non-coding sequences.
2. Analysis of low genome coverage of paired-end reads for rearrangements. Paired-end designs will be available for most new sequencing technologies. The aim of these analyses will be to identify at least 80% of somatic genomic rearrangements down to sequence level resolution. Paired-end sequence coverage should be estimated on this basis.
3. Genotyping arrays
It is recommended that a high density genotyping array be performed at an early stage on all samples in the ICGC set. This is a straightforward and inexpensive additional experiment that will provide copy number, LOH and breakpoint information that is highly useful in the interpretation of other analyses. Information from the genotyping array will also critically be useful in tracking of samples and confirming the relationship between a tumor and a normal sample.

The above studies should preferably be conducted on samples that will be entered into the final whole genome shotgun approach and therefore the data will ultimately be merged.

Class 2: Complementary genomic analyses

POLICY: Additional studies of DNA methylation and RNA expression are recommended **on the same samples that are used to find somatic mutations.**

The potential list of complementary analyses is long. The recommended supplementary analyses for ICGC have therefore been selected to be pragmatic, to have relatively easily achievable aims, to not significantly complicate sample acquisition, and to likely enrich interpretation of somatic mutation information. The ICGC members have restricted the scope of the Consortium to include only those analyses only requiring DNA and RNA.

2.1. Analyses of DNA methylation

Optimally, the outcome of DNA methylation profiling should be the assignment of the methylation state of every CpG and the identification of CpGs that are differentially methylated in the cancer versus normal genomes of the same tissues.

The current expectation is that it may not be necessary to determine the frequency/extent of DNA methylation at each and every CpG, but it may be sufficient to determine the frequency/extent of DNA methylations within a small genomic region containing multiple CpGs.

Current 'comprehensive' DNA methylation analyses have focused mostly on (CpG-island containing) promoters and readout of differential methylated regions (DMRs) on microarrays. However, DNA methylation outside CpG-islands and/or promoters may well be diagnostic and prognostic and instrumental in classification even though the expression profile in the cancer cell may not be affected by the extent of this type of DNA methylation.

A wide variety of techniques have been described for identifying/profiling of DNA methylation. They differ in the resolution of methylation mapping, the ability to give qualitative rather than quantitative measurements, and in their potential to be used in global rather than gene-specific analysis. There is consensus at this point that it is unwise to make specific recommendations regarding the methylation approach that must be used by ICGC cancer projects.

2.2. Analyses of RNA expression

There are currently several technologies and platforms for analysis of RNA expression. These continue to develop. In particular, digital quantification of expression based on sequencing technologies may become practical soon and may be optimal for these purposes. At this time, ICGC therefore is not making specific recommendations on technological approaches to be used for RNA expression studies.

It is recommended, however, that analyses of expression include all protein coding genes (or use easily available commercial platforms that include most protein coding genes) and consider some of the non-coding RNAs, notably microRNAs. Analysis of the transcriptome may be more critical in some cancer types than in others, for example in breast cancer where it is fundamental to the classification.

Class 3: Optional analyses

Although outside of the initial scope of the ICGC, other analyses of samples used in the somatic mutation screen are clearly to be encouraged. These could include:

- Proteomic analyses;
- Metabolomic analyses;
- Immunohistochemical analyses;
- Analyses of chromatin state;
- It may be particularly helpful in the long term to construct tissue arrays from the cancers in each category for future immunohistochemical and other in situ analyses.

No specific recommendation is made to ICGC members regarding approaches, platforms, and other issues related to optional analyses.

Quality control

The ability of each center to produce data of the requisite quality will be assessed by circulating a small set of inexhaustible tumor/normal samples which each center will have to analyze for each component of the project they are engaging in. It is proposed that these samples be publicly available cancer cell

lines for which a normal DNA sample is available. Cancer cell lines may be spiked with normal DNA to better recapitulate the state of a primary tumor specimen. Centers will be expected to provide coverage of these samples such that 80% of somatic alterations are detected, of which 95% should be real.

Coverage requirements for primary tumor samples analyzed by the ICGC will be estimated on an individual basis by assessment of sequence error rate in each sample and other parameters that determine sensitivity and specificity of variant detection. These ongoing quality control measures will continue to be refined by the ICGC with a view to implementation during the course of the project.

E. 9. ICGC Data Management

Similar to other large-scale genome projects, the ICGC will require a **Data Coordination Center (DCC)** that is well integrated with ICGC operations at participating centers, and ICGC Governance and Scientific Coordination bodies. This requires a comprehensive management system that is designed to:

- provide secure and reliable mechanisms for the sequencing centers, biorepositories, histopathology groups, and other ICGC participants to upload their data;
- track data sets as they are uploaded and processed, to perform basic integrity checks on those sets;
- allow regular audit of the project in order to provide high-level snapshots of the consortium's status;
- perform more sophisticated quality control checks of the data itself, such as checks that the expected sequencing coverage was achieved, or that when a somatic mutation is reported in a tumor, the sequence at the reported position differs in the matched normal tissue;
- enable the distribution of the data to the long-lived public repositories of genome-scale data, including sequence trace repositories, microarray repositories and the genome browsers;
- provide essential meta-data to each public repository that will allow the data to be understandable;
- facilitate the integration of the data with other public resources, by using widely-accepted ontologies, file formats and data models;
- manage an ICGC data portal that provides researchers with access to the contents of all franchise databases and provides project-wide search and retrieval services.

The ICGC data management system will be required to provide the following support to experimental biologists, computational biologists, and other researchers:

- support for hypothesis-driven research: The system should support small-scale queries that involve a single gene at a time, a short list of genes, a single specimen, or a short list of specimens. The system must provide researchers with an interactive system for identifying specimens of interest, finding what data sets are available for those specimens, selecting data slices across those specimens (e.g., counts of the number of somatic mutations observed a region within the UTR of a gene of interest), and running basic analytic tests on those data slices;
- support for computational biologists: The system should allow large subsets, or even the entire ICGC dataset, to be downloaded;
- enable ICGC and legislative policies for protecting the confidentiality of tissue donors, by prohibiting access to protected data to users who are not duly authorized.

Each data producer will manage its own workflow and be responsible for primary QC, data integrity and protection of confidential information. A common core of ICGC data intended for integration and redistribution will be shared with the research community via local “franchise databases” which share a common data model and structure. The franchise database software (schema, integrity-checking utilities, load and dump utilities) will be written by the DCC and managed by the data producers. Under this architecture, ICGC participants can develop their own project-specific data models, workflows, and databases. At regular intervals, a subset of the information contained in the project-specific databases will be exported into a local ICGC franchise database, which will implement a uniform simplified data model that captures the essential data elements that are needed to implement ICGC-wide policies on data release, quality control and milestones. The franchise will also include a set of standardized validation and quality control tools, developed and deployed by the ICGC coordinating body, that are used to validate that the information placed in the franchise database is complete and internally consistent.

In order to provide the research community with a single portal into the entire ICGC data set, a coordination backend database will act as the union of all the franchise databases. The coordination backend will use the same data model as the individual franchise databases, but will appear to users as though it contains all the ICGC data in one place. This effect can be achieved either via a physical mirroring process in which the coordination backend pulls in copies of each of the franchise databases at regular intervals, or via a passthrough system in which queries directed at the coordination backend are multiplexed among the individual franchise databases.

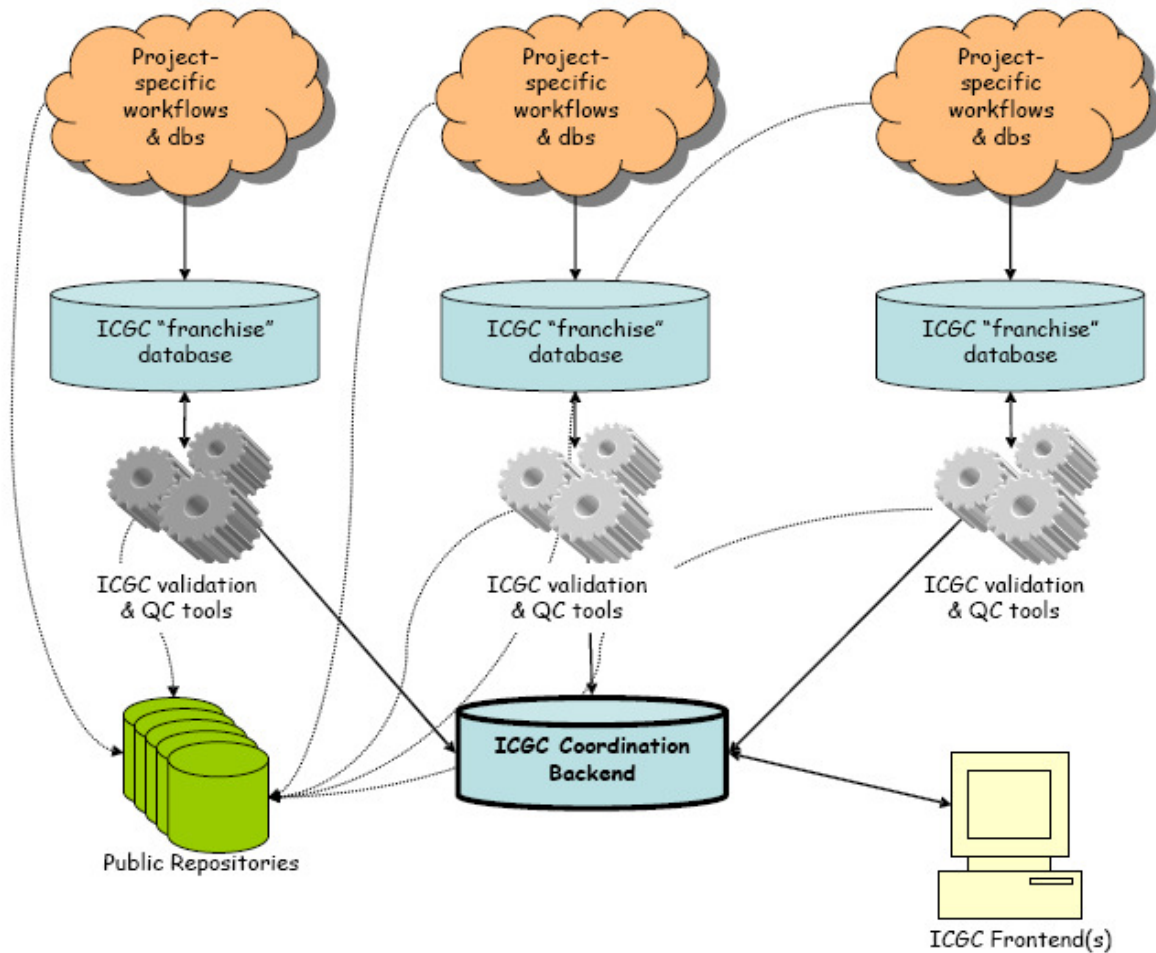


Figure 2: ICGC data coordination as a franchise system

The community will obtain access to the ICGC data via one or more front ends (e.g., websites), that will provide an interface to the coordination backend. In addition, all project data will be submitted to the appropriate public repositories. The exact path that the data will take from the group that generates it to the public repository will be flexible. For some data types it would be appropriate for the ICGC participant to submit the information directly from their internal workflow system. In other cases, it might be appropriate for the information to be submitted from the franchise database or from the coordination database itself. This architecture provides the flexibility to allow certain specialized ICGC data types - microarray CEL files, raw short sequence reads, details of tumor-specific treatment regimens, histology slide images - to be submitted directly to the appropriate archive without bottleneaking through a central coordinating center or generic data model. Nevertheless, by whichever path the detailed data takes to the repository, the tracking information needed to connect the sample data to that detailed information will be captured by the franchise database and available to researchers via the coordinating back end.

Box 8 includes additional recommendations with respect to requirements for data storage, analysis, distribution and protection.

Box 8. Additional guidelines for ICGC data management and security

Quality standards: Periodic quality assurance exercises, such as round-robin validation experiments, should be coordinated and interpreted by the DCC. The results of these validation exercises will be made available via the ICGC data portal.

Public and protected tiers: A binary system shall apply to portions of the data such that a datum is either *public*, meaning that all end-users can gain access to it, or *protected*, meaning that access is only available to authorized researchers who have agreed to protect patient confidentiality.

Multilateral authorization: The ICGC should have multiple bodies that can authorize a researcher to gain access to protected data as per IDAC Policies. Once a researcher is authorized by any of these bodies, he or she should be granted access to all protected ICGC data, regardless of which collaborator generated it or which country the data resides in.

Other portals: The ICGC should encourage the redistribution, integration and visualization of the data by community bioinformatics portals. However, portals that provide access to protected data sets must agree to respect and to implement ICGC's authentication and authorization standards for protection of patient confidentiality.

Submission to archival repositories: The unprotected portion of the data should be submitted to public data repositories as rapidly as possible after passing QC and other verification tests.

Use of community standards: Whenever possible, the ICGC coordinating center and participating data acquisition groups should represent data sets using existing community file formats, ontologies and other standards.

Analysis services: Analysis and data aggregation services, which may be deployed against the ICGC data sets, will sometimes need to be co-located with the primary data in order to provide acceptable performance. In the event that a primary data set resides in a public archive, such as the short read archive, this will require coordination between the ICGC and the archive managers.

**APPENDIX: PARTICIPANTS IN THE PLANNING PHASE OF THE ICGC
(NOV 2007 – MARCH 2008)**

INTERNATIONAL CANCER GENOME CONSORTIUM (ICGC) WORKING GROUPS						
Clinical and Pathology Issues	Quality Standards of Samples	Genome Analyses	Informed Consent and Privacy Protections	Sample Size/Study Design	Data Management/ Databases and Coordination	Data Release, Data Tiers, Intellectual Property, and Publications
Lynda Chin Jean-Yves Blay William Dalton Tony Green Stan Hamilton Timothy Ley Ed Liu Paul Mischel Kenneth Pienta Rajiv Sarin Daniel Tan	Peter Lichter Carolyn Compton Andy Futreal Youyong Lu Miguel Angel Piris	Mike Stratton Olli Kallioniemi Ed Liu Marco Marra John McPherson Brad Ozenberger Henk Stunnenberg Daniel Tan Brandon Wainwright Rick Wilson	Bartha Knoppers Martin Bobrow Wylie Burke Kazuto Kato Karen Kennedy Brad Ozenberger Daniel Tan Susan Wallace Henry Yang	Eric Lander Ron DePinho Doug Easton Gaddy Getz Partha P. Majumder	Lincoln Stein Cameron Brennan Arul Chinnaiyan Peter Good Joe Gray J Gowrishankar David Haussler David Housman Tim Hubbard Subha Madhavan Paul Spellman	Mark Guyer Daniela Gerhard Karen Kennedy Brad Ozenberger

ICGC SCIENTIFIC PLANNING COMMITTEE (SPC)				
Warwick Anderson	Ron DePinho	Tim Hubbard	Edison Liu	Lincoln Stein
Anna Barker	Doug Easton	Tom Hudson	Partha P. Majumder	Mike Stratton
Cindy Bell	Andy Futreal	Olli Kallioniemi	Marco Marra	Henk Stunnenberg
Martin Bobrow	Daniela Gerhard	Karen Kennedy	Brad Ozenberger	Brandon Wainwright
Wylie Burke	Tony Green	Bartha Knoppers	Jane Peterson	Rick Wilson
Francis Collins	Mark Guyer	Eric Lander	Alan Schafer	Henry Yang
Carolyn Compton	Stan Hamilton	Timothy Ley	Paul Spellman	(represented by Prof. Youyong Lu)
William Dalton		Peter Lichter		

INTERIM ICGC EXECUTIVE (EXEC)
Warwick Anderson, National Health and Medical Research Council, Australia (Observer Status)
Cindy Bell and Karen Kennedy, Genome Canada, Canada (Observer Status)
Tom Hudson, Ontario Institute for Cancer Research, Canada
Henry Yang, Chinese Cancer Genome Consortium, China
Jacques Remacle, Patrik Kolar and Iiro Eerola, European Commission (Observer Status)
M.K. Bhan and T.S. Rao, Ministry of Science & Technology, Department of Biotechnology, India
Edison Liu, Genome Institute of Singapore, Singapore
Alan Schafer and Michael Stratton, The Wellcome Trust; Wellcome Trust Sanger Institute, United Kingdom
Anna Barker and Daniela Gerhard, National Cancer Institute, United States
Francis Collins, Jane Peterson, Mark Guyer and Brad Ozenberger, National Human Genome Research Institute, United States

The ICGC Executive and Scientific Planning Committee thank Jennifer Jennings at the Ontario Institute for Cancer Research for her invaluable assistance in coordinating teleconferences, minutes, reports and other communications.

