Science-Metrix

Bibliometric Assessment of Research Funded by Genome Canada 1996–2007

Science-Metrix

Bibliometric Assessment of Research Funded by Genome Canada 1996–2007

January 13, 2009

by

David Campbell, M.Sc.

Isabelle Labrosse, M.Sc.

Grégoire Côté, B.Sc. and

Éric Archambault, Ph.D.

submitted to

Genome Canada

514.495.6505 ▪ 1335 A Mont-Royal E. ▪ Montréal ▪ Québec ▪ Canada ▪ H2J 1Y6

info@science-metrix.com ▪ www.science-metrix.com ▪ www.rd-reports.com

Executive Summary

Genome Canada was established in April 2000 by the federal government to provide Canada with the capacity to undertake large-scale projects that would secure its position as a world leader in genomics in areas of strategic importance to Canadians (e.g., agriculture, environment, fisheries and health). Being a publicly funded corporation, Genome Canada is accountable for the value it creates through the funding of research and is therefore committed to the highest standards of good governance and responsible management. It is in this context that Genome Canada is undergoing a second formative evaluation. As part of this exercise, Science-Metrix was mandated to provide performance measurements of the large-scale research projects supported by Genome Canada under its third objective, that is, to support the most promising genomics projects to be performed by outstanding researchers that could not have been funded through existing mechanisms given their scope and scale.

The present study's key findings are:

- The papers produced by researchers funded by Genome Canada had significantly higher observed and expected scientific impacts than other genomics papers from Canada and the world. This suggests that Genome Canada successfully implemented its peer-review process for the selection of outstanding researchers.
- The scientific production of Genome Canada's supported researchers seems to have increased more rapidly while they were receiving funds from Genome Canada than while they were not.
- The production of genomics papers increased for nearly 65% of the researchers with support from Genome Canada, and the difference in their production before and while they were funded by Genome Canada is statistically significant.

  However, care should be taken in interpreting these results in terms of determining the direct effect of Genome Canada's funding, as they may be caused by other factors. For example, if Genome Canada supported a number of young and promising researchers whose scientific production usually exhibit exponential growth early in their career, this pattern may have been observed when analyzing the same group of researchers without Genome Canada's support.
- In testing the effect of Genome Canada funding on the scientific impact of research, two statistical tests were performed. The first was a test for two independent samples comparing the overall scientific impact of Genome Canada's supported versus non-supported papers; it was found that papers authored with financial support had significantly higher impact than non-supported papers. The second test was for two related samples, comparing the scientific impact of individual researchers before and during the period of support by Genome Canada. This test found no significant difference; in fact, the scientific impact increased for slightly less than half of the researchers. As such, not all researchers with financial support from Genome Canada contributed to the higher impact of supported versus non-supported papers.
- Even if the difference between the impact of the supported and non-supported papers is statistically significant, it remains modest and can be attributed to a small percentage of high impact papers.
- Among the high-impact papers are success stories that would likely not have been accomplished had Genome Canada not been created. These success stories undoubtedly contributed to the prestige of Canada and its recognition as a world leader by its international peers.
- The substantial financial resources made available through Genome Canada for funding large-scale genomics projects appear to have been a factor in attracting outstanding scientists to Canada.
- It is not surprising that funding the best researchers produces great results, but it does not change the research landscape profoundly nor does it fuel the achievement of astonishingly greater scientific excellence. As such, it might be pertinent to attribute part of the funds to teams composed of promising researchers (i.e., with innovative projects) rather than to outstanding researchers. This is surely more risky and selection models for these researchers are few and far between, but it might be one of the most potent ways to inject vitality in the research system.

Contents

Executive Summary ..... **i**
Contents ..... **ii**
Tables ..... **ii**
Figures ..... **ii**
1 Introduction **3**
2 Methods **5**
2.1 Database 5
2.2 Constitution of datasets 5
2.2.1 Constitution of the genomics dataset 6
2.2.2 Constitution of the dataset of Genome Canada's funded applicants 6
2.3 Bibliometric Indicators 9
2.4 Statistical Analyses 11
3 Results & Discussion **13**
3.1 Genome Canada's supported researchers compared to Canadian and World Researchers 13
3.2 Production of Genome Canada's supported researchers while they were receiving funds from Genome Canada and while they were not 14
3.3 Scientific impact of Genome Canada's supported researchers while they were receiving funds from Genome Canada and while they were not 16
3.4 Success stories—highly cited papers supported by Genome Canada 18
3.5 Attracting outstanding scientists to Canada 20
Conclusion **21**

Tables

Table I Comparison of the observed (ARC) and expected (ARIF) scientific impact of the papers produced by researchers funded by Genome Canada (including supported and non-supported papers) with those of other genomics papers from Canada and the world, 1996–2007 13
Table II Comparison of the observed (ARC) and expected (ARIF) scientific impact of Genome Canada researchers' supported and non-supported papers in the genomics dataset, 1996–2007 16
Table III Proportion of Genome Canada's supported and non-supported papers within quartiles of the citation distribution of world genomics papers and within the 5% most cited genomics papers in the world, 1996–2005 17
Table IV Proportion of Genome Canada's supported and non-supported papers within quartiles of the distribution of world genomics papers' impact factors and within the 5% of world genomics papers with the highest IF, 1996–2007 17
Table V The top 10 papers in terms of citation scores among the papers authored with support from Genome Canada, 2003–2005 19
Table VI Percentage of genomics papers with a Canadian address among non-supported and supported genomics papers, 1996–2007 20

Figures

Figure 1 Trends in the number of published papers per researcher for a group of researchers who received support from Genome Canada, up to 11 years prior to receiving support and up to 5 years with support, 1996–2007 14
Figure 2 Trend in the percentage of genomics papers by researchers who received support from Genome Canada, up to 11 years prior to receiving support and up to 5 years with support, 1996–2007 16
Figure 3 Observed (ARC) and expected (ARIF) scientific impact of supported papers 18

1 Introduction

Since the establishment of the Human Genome Project in the late 1980s, governing bodies around the world have invested substantial financial resources into the creation of an International Human Genome Sequencing Consortium. This Consortium is made up of genomics centres with high production capacities, the foremost of which are located in the US (the Whitehead Institute for Biomedical Research and the Centre for Genome Research) and the UK (the Wellcome Trust Sanger Institute). Although Canada was among the leading countries in genome science (ranking 6th in 1996–1998) and technology (ranking 3rd in 1996–1998) in the mid 1990s,[1] a high-throughput genome centre was not established in Canada during that period, preventing the country from taking part in the Human Genome Project. Project participants from the countries representing the five largest producers of genomics papers (the US, Japan, the UK, Germany and France), as well as from China, eventually captured the human genome sequence, which was subsequently published in a 2001 issue of *Nature*.[2] In April 2000, the federal government created Genome Canada to confer to Canada the capacity to participate in such large-scale projects and to secure its position as a world leader in genomics in areas of strategic importance to Canadians (e.g., agriculture, environment, fisheries and health).

Genome Canada undertook its mandate based on a business-oriented model, one that would target the support of substantially larger research projects than are usually funded by other Canadian funding agencies; the creation of genome centres with the required support infrastructure, and; a strong emphasis on research management by these centres.[3] To date, Genome Canada has invested $840 million to meet nine underlying objectives:

1. Promote interaction between industry, governments, universities, hospitals, research institutes, and the public in realizing the goals of the national genomics initiative.
2. Set up six genome centres (Vancouver, Calgary, Saskatoon, Toronto, Montreal and Halifax) providing access to leading-edge technologies to researchers across Canada, and support the training of highly qualified workforce in genomics.
3. Support the most promising genomics projects to be performed by outstanding researchers that could not have been funded through existing mechanisms given their scope and scale.
4. Support the establishment of the necessary science and technology platforms to carry out the large-scale projects (e.g., high-throughput platforms for functional genomics and proteomics, genomics sequencing, genotyping, bio-informatics, and new technology development).
5. Lead scientific research in the ethical, environmental, economic, legal, and social issues related to genomics (GE^3LS).
6. Communicate the results of genomics research to the public.
7. Enable and promote Canadian participation in international genomics research programs.

---

[1] Campbell, D., Côté, G. and Archambault, É. 2008. Benchmarking of Canadian Genomics — 1996–2007. Prepared for Genome Canada, 13 pages.

[2] International Human Genome Sequencing Consortium. 2001. Initial Sequencing and Analysis of the Human Genome. Nature, 409(15): 860-921.

[3] BearingPoint. 2004. Interim Evaluation of Genome Canada. Prepared for Genome, 53 pages.

8. Foster investment by others through requirements for matching funds.
9. Pursue and ultimately achieve economic, industrial and social benefits to Canada.

Being a publicly funded corporation, Genome Canada has stated its commitment to *"upholding the highest standards of good governance and responsible management in carrying out its mandate*".[4] As such, Genome Canada has set forth a Performance, Audit and Evaluation Strategy (PAES) to ensure accountability in the achievement of its objectives.[5] In 2003–2004, the organization underwent a first formative evaluation, and it is currently being evaluated for the second time. Within the context of this second evaluation, Science-Metrix has been mandated to provide performance measurements of the large-scale research projects supported by Genome Canada (see above description of objective 3). Specifically, the work performed by Science-Metrix will be instrumental in:

- determining whether or not Genome Canada was successful in selecting outstanding researchers by comparing the scientific performance of the researchers it supported to that of the average of Canadian and world researchers (excluding Genome Canada's supported researchers) in genomics (Section 3.1);
- comparing the scientific production (Section 3.2) and impact (Section 3.3) of Genome Canada's supported researchers while they were receiving funds from Genome Canada and while they were not;
- presenting success stories, identified by selecting the most cited papers supported by Genome Canada (Section 3.4); and
- providing evidence as to whether or not Genome Canada has attracted outstanding researchers from abroad (Section 3.5).

It should be noted that this bibliometric assessment is not designed to provide evidence regarding the broader impact Genome Canada might have had on the country's overall scientific production through the establishment of genome centres aimed, in part, at making accessible leading-edge technologies to the Canadian community of researchers pursuing genomics research (see above description of objective 2).

[4] See http://www.genomecanada.ca/.

[5] See http://www.genomecanada.ca/medias/PDF/EN/PerformanceAuditandEvaluationStrategy.pdf.

2 Methods

The selection of the bibliographic database for the constitution of the datasets used in producing reliable indicators of scientific production for Genome Canada's funded researchers is discussed in Section 2.1. Section 2.2 addresses the construction of these datasets in detail, while Section 2.3 presents the bibliometric indicators used to quantify scientific outputs. Finally, Section 2.4 describes the statistical analysis performed on bibliometric indicators.

2.1 Database

Access to a database containing the most complete bibliographic information on scientific serials published worldwide is essential for the gathering of bibliometric data. In this study, Thomson Reuters' Web of Science (WoS), which includes three databases (the *Science Citation Index Expanded™* [SCI Expanded], the *Social Sciences Citation Index™*, and the *Arts & Humanities Citation Index™*) covering the complete spectrum of scientific fields (e.g., natural sciences and engineering [NSE], social sciences and humanities [SSH]), was used to produce statistics on the scientific production of Genome Canada's supported researchers.

The WoS was chosen because it indexes some 9,000 of the world's most cited refereed journals (i.e. about 1,500,000 peer-reviewed scientific documents each year), which are generally regarded by the scientific community as the most renowned and reliable journals available in their respective fields. Furthermore, unlike Medline, the WoS lists the cited references of each document it includes (e.g., articles, chapters published in journals or book series). This permits the analysis of the scientific impact of publications based on citation counts and the impact factor.[6]

Although the WoS lists several types of documents, only articles, research notes, and review articles were retained in the production of the bibliometric indicators, as these are considered to be the main types of documents through which new knowledge is disseminated in the NSE. In addition, these documents have been subject to peer review prior to being accepted for publication, ensuring that the research is of good quality and constitutes an original and robust contribution to scientific knowledge. In this report, articles, notes, and reviews are collectively referred to as "papers".

2.2 Constitution of datasets

This section details how the genomics dataset (Section 2.2.1) and the dataset of Genome Canada's funded applicants (Section 2.2.2) were produced.

[6] See: http://scientific.thomsonreuters.com/free/essays/journalcitationreports/impactfactor.

2.2.1 Constitution of the genomics dataset

The dataset was constructed by querying genome-specific keywords in the titles of papers indexed in the WoS. The keyword set for the query was originally defined in 1999 by experts appointed by Genome Canada and by Science-Metrix analysts. This keyword set was extensively revised (in 2000 and again in 2003) to remove as many false positives as possible and to achieve maximum coverage. The resulting datasets comprise papers in core and peripheral genomics (such as papers in molecular biology that touch upon genome research). Some keywords were used only within the boundary of genomics (e.g., the term "sequenc*" can apply, for instance, to mathematics; therefore, these papers were excluded). The resulting keyword set is listed below (the * represents a wildcard character, which means that any word starting with the letters preceding the * were included).

2-hybrid	Genom*	Mutation	Proteom*
Allel*	Genot*	Nucleic acid	QTL
Antisense	Haplotyp*	Nucleosid	Radiation hybrid map
Autosom*	Intron	Nucleotid*	Recombinant
Biochip	Linkage map	PCR	Restriction map
Candidate region	Loci	Physical map	RNA (mRNA, RNAi, etc.)
Chromoso*	Locus	Plasmid*	Sequenc*
Cloning	Meiosis	Ploidy	SNP
DNA (cDNA, anti-DNA, etc.)	Meiotic	PNA	Transcript*
Exon	Microsatellite	Polymerase	Transgen*
Expressed	Minisatellite	Post-translation	YAC
Genetic*	Molecular characterization		

2.2.2 Constitution of the dataset of Genome Canada's funded applicants

A bibliometric dataset for an institution is usually built by retrieving papers in which the name of the institution is found in the authors' address. Because Genome Canada is an organisation that supports research as opposed to a research institute *per se*, its name is not expected to be found in the address field of papers published by the researchers it funds. This makes it virtually impossible to precisely identify those papers that were produced with financial support from Genome Canada. As a result, to build a dataset of papers supported by Genome Canada, a publication portfolio was built for each of the principal investigators (PI) who received funding from Genome Canada within the context of its third objective (i.e., support the most promising genomics projects, to be performed by outstanding researchers, which could not have been funded through existing mechanisms given their scope and scale).

Unfortunately, this approach will not capture all of the papers of all researchers that might have benefited from accessing the leading-edge technologies made available through the genome centres established by Genome Canada (see objective 2, Section 1). As such, even if no tangible evidence of the positive effect of Genome Canada is found, it will not be possible to conclude that Genome Canada has not had a broader impact on the scientific performance of the Canadian community of researchers pursuing genomics research.

To build the dataset, Science-Metrix first obtained a listing, provided by Genome Canada, of PIs who were awarded grants to pursue large-scale projects in Competition I (2001–2002), Competition II

(2002–2003), the Canada/Spain Competition (2004–2005), the Applied Human Health Competition (2004–2005), and Competition III (2005–2006).

The names of these researchers were then used in an automatic query that retrieved their WoS-indexed scientific output. Before executing the automatic retrieval of PIs' papers, the names as they appear in Genome Canada's list were transformed to match the format of author names in the WoS. Author names in the WoS do not include authors' first names, only their initials. For example, "John W. Smith" is transformed into "Smith-JW", as well as "Smith-J". The latter form ensures that publications for which the middle name (or its initial) is omitted are retrieved. Subsequently, the formatted names are queried against the database to retrieve, for each researcher, all of the papers bearing his/her name as an author between 1996 and 2007. The search was first limited to papers bearing a Canadian address to minimize the occurrence of false positives (i.e., papers belonging to another researcher with the same surname and initials) resulting from homographs in researchers' names. However, to ensure that all papers by an author were retrieved, additional searches were performed against the database focusing on the institutions and countries visited by the researchers, as revealed by their educational and professional backgrounds. Nevertheless, these queries inevitably overestimated the number of publications in many paper portfolios, especially for researchers with a common surname (e.g., Smith). Since there is no *a priori* regarding which researchers will be overestimated and which will not, the papers that are retrieved automatically must be validated manually for each researcher to remove false positives.

In so doing, careful attention was paid to the disciplines and specific topics of papers belonging to a publication portfolio. Several questions arise when analysing whether or not a set of papers belong to a given researcher (e.g., Are those papers consistent with respect to the discipline of the researcher as revealed by his/her departmental affiliation? Is the scope of those papers broader than the products of only one individual researcher?). For example, the attribution of an engineering paper to a biologist or a physics paper to an historian would be seriously questioned. However, given the commonness of multidisciplinarity in science, it is not sufficient to rely mechanistically on departmental affiliations of researchers to validate the publications of their portfolio; a philosopher may publish articles dealing with medical ethics in clinical medicine journals, and an engineer may collaborate on papers dealing with environmental problems published in biology or earth sciences journals. The institutional addresses may provide additional clues, since they often include the authors' departments (although these are not harmonized in the WoS).

In cases where the previous actions failed to determine whether a paper is part of a researcher's portfolio, the publication was downloaded when it was electronically available through libraries or open access. The signatures on the originating paper often provide a link between each author's name and his or her institutional address (including departmental affiliation), which normally allows one to unambiguously identify false positives. Altogether, manual cleaning of publication portfolios is a time- and resource-consuming process requiring careful attention. Nonetheless, it is currently the only way to guarantee that results are sufficiently robust to evaluate important questions, such as the impact of funding on specific groups of researchers.

In addition to false positives, false *negatives*, or papers authored by a researcher that were not retrieved by the automatic query, are a concern. These "absent papers" reflect the fact that the WoS only covers a fraction of all work published worldwide by researchers. For example, journals of national interest, books, and various official publications that are generally referred to as "grey literature" (including minutes from conferences and symposiums, research reports, and in-house journals) are not indexed in Thomson Reuters' scientific databases. Therefore, the publications in the WoS do not encompass the entire CV of researchers funded by Genome Canada. Nevertheless, the WoS indexes the portion of their publications that is the most visible and the most frequently cited by the scientific community.

In total, 81 projects were selected based on the period during which financial support was provided by Genome Canada (the beginning and ending fiscal years) to allow for the occurrence of supported papers. Papers were considered to be "supported" if they were published in the period starting two years after the first year of financial support and ending two years after the last year of financial support. As such, a project funded between 2006/2007 and 2009/2010 would not be included, as only papers published between 2008 and 2011 would be considered supported, and the study covers only papers published between 1996 and 2007. Papers published between 1996 and 2007 but which fall outside the period during which papers are considered as being supported constitute "non-supported" papers. Given the method used to classify papers, the "supported papers" may include, in addition to articles published as a result of projects funded by Genome Canada, papers authored during the period of Genome Canada support but as a result of other projects funded by other agencies.

To maximize the sensitivity of the analysis, it would have been necessary to tag non-supported papers as being published before or after the period of support. Unfortunately, not enough time has elapsed since Genome Canada's first competition (2001) to allow for an analysis of non-supported papers published after the period of support. The first projects were funded in 2001 and most of them were funded for at least three years, such that the first non-supported papers following support could only have been published starting in 2005. To date, the period under which papers are considered to be supported has ended for only one project, such that there are currently only two non-supported papers published after the period of support. Consequently, the analyses presented in this report only compares papers published before the period of support (non-supported papers) to papers published during the period of support (supported papers).

As the focus of the current study is on genomics, most of the analyses were performed only on the portion of papers by Genome Canada's supported researchers that intersect with the genomics dataset, thereby removing nearly half of the production of these researchers. Within the portion of their scientific production not classified in genomics are papers not related to the field and false negatives (i.e., genomics papers that were not classified in genomics using keyword searches limited to the titles of papers). These false negatives do not create biases in the analysis, as they occur at the same frequency within supported and non-supported papers (estimated based on a random sampling of 30 supported and 30 non-supported papers outside of the genomics dataset; Z-test for two proportions; data not shown).

2.3 Bibliometric Indicators

Using researcher portfolios built using the aforementioned methods as well as papers computed at the world and country (i.e., Canada) levels, the following indicators were calculated:

Number of publications: A count of the number of scientific papers written by authors associated with a funding organization (i.e., Genome Canada) based on author names or with a country based on author addresses.

To establish a trend in the average production of Genome Canada's supported researchers (i.e., average number of papers per researcher per year) before and during the period of support, the papers were reclassified based on the number of years before support and while under support; year 0 corresponds to the first year under which papers are considered to be supported based on the approach described in section 2.2.2. This allowed data from projects having different time windows (i.e., covering different years and/or of different lengths) to be pooled. For example, a 2003 paper was tagged as supported for a project funded between 2001 and 2005 but was tagged as non-supported for a project funded between 2003 and 2007. In the first case, the paper is reclassified as year 0 whereas in the second case it is reclassified as year -2. Given the period covered (1996–2007) and the time windows of selected projects, the trend extended from 11 years prior to receiving support (-11) to 5 years with support (+4). For each researcher, the range of reclassified years (i.e., between -11 and +4) during which they could have published papers, based on the time window of their respective projects, was determined. This was used to calculate the average number of papers per researcher per reclassified year.

Average of Relative Citations (ARC): This is an indicator of the *observed* scientific impact of papers produced by a given entity (e.g., a country, a specific set of papers, a researcher) that considers citations in papers published in peer-reviewed journals. The total number of citations received for each paper was counted such that the citation window of papers published in different years differ (i.e., a 1996 paper accumulated citations over 12 years, whereas a 2005 paper accumulated citations over 3 years). To account for this variation in the citation windows of papers and for the different citation patterns across fields and subfields of science (e.g., there are more citations in biomedical research than mathematics), the citation count of a paper in a given subfield (the classification of papers by subfield is based on the journal classification of the US National Science Foundation[7]) is divided by the average count of all papers (within the WoS) published the same year in its subfield, to obtain a relative citation count (RC). The ARC of a given entity (e.g., a country, a specific set of papers, or a researcher) is the average of the RC of papers belonging to it.

In general, health research papers reach their citation peak (year in which they have received the most citations) about two to three years after publication.[8] Therefore, many articles published in 2006 and

[7] See: http://www.nsf.gov/statistics/seind06/.

[8] See: http://www.in-cites.com/ESI_Product_Info/1-HotPapers.htm.

2007 will not have reached their citation peak, leading to misleading citation scores by subfield and year (i.e., average of citations for articles in a subfield in 2006 or 2007 might be biased downward), such that their inclusion can render the ARC indicator unreliable when used at a 'micro' level (i.e., small groups of researchers). As a result, only papers published between 1996 and 2005 were included when computing the ARC indicator. The complementary use of the ARIF indicator (expected impact; see below) is useful in that it permits the analysis of scientific impact up to the most recent year analyzed.

Number of most cited papers: The number of papers that are in the 5% of papers with the highest RC (see above definition in the ARC's description), for a given an entity.

Average of relative impact factors (ARIF): This is an indicator of the *expected* scientific impact of papers produced by a given entity (e.g., a country, a specific set of papers, a researcher) based on the journals in which they were published. Thomson Reuters calculates an annual impact factor (IF) for each journal based on the number of citations it received in the previous two years, relative to the number of papers it published in the previous two years. Thus, each journal's IF will vary from year to year. The IF of a journal in 2007 is equal to the number of citations to articles published in 2006 (8) and 2005 (15) divided by the number of articles published in 2006 (15) and 2005 (23) (i.e., IF = numerator [23] / denominator [38] = 0.605).[9] However, as pointed out by Moed and colleagues (1999), Thomson Reuters' IF is flawed in that its numerator and denominator are not symmetric:

> ISI classifies documents into types. In calculating the nominator of the IF, ISI counts citations to all types of documents, while as citable documents in the denominator ISI includes as a standard only normal articles, notes and reviews. However, editorials, letters, and several other types are cited rather frequently in a number of journals. When they are cited, these types do contribute to the citation counts in the IF's numerator, but are not included in the denominator. In a sense, the citations to these documents are "for free".[10]

In this study, we therefore used a symmetric IF based on three documents types (i.e., articles, notes, and reviews) which is computed by the Observatoire des Sciences et des Technologies (OST) du Québec using Thomson Reuters' WoS.

The IF of papers is calculated by ascribing to them the IF of the journals in which they are published, for the year in which they are published. Subsequently, to account for different citation patterns across fields and subfields of science (e.g., there are more citations in biomedical research than mathematics), each paper's IF was divided by the average IF of the papers published the same year in its subfield to obtain the Relative Impact Factor (RIF). The ARIF of a given entity is the average of its RIFs (i.e., if an institution has 20 papers, the ARIF is the average of 20 RIFs, one per paper).

Number of papers in highest impact journals: The number of papers that are in the 5% of papers with the highest RIF (see above definition in the ARIF's description), for a given entity

---

[9] See: http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/.

[10] Moed, H.F., Van Leeuwen, T.H.N., Reedijk, J. 1999. Towards appropriate indicators of journal impact. *Scientometrics*, 46(3): 575-589.

2.4 Statistical Analyses

To establish whether there were significant differences between various entities in terms of scientific production and scientific impact, a series of statistical tests were performed in SPSS. For each statistical test, the difference was considered to be significant at $p < 0.05$.

Because data on scientific production and scientific impact are not normally distributed, non-parametric tests were used. Here is a list of null hypotheses that were tested for significance. Except when specified otherwise (in brackets), the Mann-Whitney U test was used:

- H_o = The ARC of genomics papers authored by researchers who were supported by Genome Canada (including supported and non-supported papers) is not significantly different from the ARC of Canada's genomics papers (excluding papers by Genome Canada researchers);
- H_o = The ARC of non-supported genomics papers by Genome Canada's supported researchers is not significantly different from the ARC of Canada's genomics papers (excluding papers by Genome Canada researchers);
- H_o = The ARC of genomics papers authored by researchers who were supported by Genome Canada (including supported and non-supported papers) is not significantly different from the ARC of world genomics papers (excluding papers by Genome Canada researchers);
- H_o = The ARC of non-supported genomics papers by Genome Canada's supported researchers is not significantly different from the ARC of world genomics papers (excluding papers by Genome Canada researchers);
- H_o = The ARIF of genomics papers authored by researchers who were supported by Genome Canada (including supported and non-supported papers) is not significantly different from the ARIF of Canada's genomics papers (excluding papers by Genome Canada researchers);
- H_o = The ARIF of non-supported genomics papers by Genome Canada's supported researchers is not significantly different from the ARIF of Canada's genomics papers (excluding papers by Genome Canada researchers);
- H_o = The ARIF of genomics papers authored by researchers who were supported by Genome Canada (including supported and non-supported papers) is not significantly different from the ARIF of world genomics papers (excluding papers by Genome Canada researchers);
- H_o = The ARIF of non-supported genomics papers by Genome Canada's supported researchers is not significantly different from the ARIF of world genomics papers (excluding papers by Genome Canada researchers);
- H_o = The yearly production of genomics papers produced by Genome Canada's supported researchers while they were receiving funds from the agency is not significantly different from when they were not funded [Wilcoxon signed rank test for two related samples];
- H_o = The ARC of Genome Canada's supported papers is not significantly different from the ARC of non-supported papers in genomics;
- H_o = The ARIF of Genome Canada's supported papers is not significantly different from the ARIF of non-supported papers in genomics;
- H_o = The ARC of Genome Canada's supported researchers in genomics while they were receiving funds from the agency is not significantly different from when they were not funded [Wilcoxon signed rank test for two related samples];

- H_o = The ARIF of Genome Canada's supported researchers in genomics while they were receiving funds from the agency is not significantly different from when they were not funded [Wilcoxon signed rank test for two related samples];
- H_o = The proportion of genomics papers that are in the 5% most cited genomics papers in the WoS is not significantly different between supported and non-supported papers [Z-test for two proportions]. The same hypothesis was tested for each quartile of the citation distribution of world genomics papers;
- H_o = The proportion of genomics papers that are in the 5% of papers with the highest impact factors in the WoS is not significantly different between supported and non-supported papers [Z-test for two proportions]. The same hypothesis was tested for each quartile of the distribution of world genomics papers' impact factors;
- H_o = The proportion of genomics papers with a Canadian address is not significantly different between supported and non-supported papers [Z-test for two proportions].

3 Results & Discussion

As a first step, the current study aimed to assess Genome Canada's success in selecting the most promising researchers to conduct large-scale genomics projects (assessing the execution of objective 3; Section 3.1). Subsequently, it examined the effect of Genome Canada's funding on the performance of supported researchers by comparing their output while they were receiving funds from Genome Canada and while they were not (performance measurement of funded research under objective 3; Sections 3.2, 3.3, 3.4, and 3.5). In the following sections, the term "supported papers" refers to scientific articles published by researchers during the period while they were receiving funds from Genome Canada. These "supported papers" may therefore include, in addition to articles published as a result of projects funded by Genome Canada, papers authored during the period of Genome Canada support but as a result of other projects funded by other agencies.

3.1 Genome Canada's supported researchers compared to Canadian and World Researchers

As the selection of outstanding researchers is a crucial component of Genome Canada's third objective, Science-Metrix compared the scientific impact of Genome Canada's supported researchers to that of the average of Canadian and world researchers (excluding Genome Canada's supported researchers). This comparison will offer evidence as to whether or not Genome Canada's selection process has been successful. If so, supported researchers should have achieved higher scientific impact.

Two indicators of scientific impact were computed: the average of relative citations (ARC), which measures the observed impact of scientific research, and the average of relative impact factors (ARIF), which measures its expected impact (Table I).

Table I Comparison of the observed (ARC) and expected (ARIF) scientific impact of the papers produced by researchers funded by Genome Canada (including supported and non-supported papers) with those of other genomics papers from Canada and the world, 1996–2007

Group	ARC*	ARIF
Genome Canada Researchers	1.82	1.50
Canada without GC Researchers	1.24	1.22
p-value	< 0.01	< 0.01
Genome Canada Researchers	1.82	1.50
World without GC	1.15	1.15
p-value	< 0.01	< 0.01
Canada	1.28	1.25
World	1.15	1.15

Note: * Relative citation counts are unreliable at this analytical level for the most recent years (2006 and 2007); therefore, papers published in those years were not included in computation of the ARC indicator (see Section 2.3).

Source: Calculated by Science-Metrix using the WoS

Because the ARC indicator is unreliable in recent years (i.e., 2006 and 2007) when used at a 'micro' level (i.e., small groups of researchers), the complementary use of the ARIF indicator is useful in that it permits the analysis of scientific impact up to the most recent year analyzed.

Results based on these two indicators support preliminary findings of the interim evaluation of Genome Canada, which found that the funding agency's selection process was successful in identifying highly promising researchers to pursue large-scale genomics projects.[3] Indeed, the papers produced by researchers funded by Genome Canada (including supported and non-supported papers) had significantly higher observed and expected scientific impacts than other genomics papers from Canada and the world for the 1996–2007 period (Table I). This finding remains true even if only non-supported papers (i.e., papers produced by Genome Canada's supported researchers while they were not receiving funds from Genome Canada) are considered (data not shown).

3.2 Production of Genome Canada's supported researchers while they were receiving funds from Genome Canada and while they were not

To assess the effect of Genome Canada's financial support on the scientific production of the researchers it supported, the average number of papers produced per funded researcher was examined up to 11 years prior to receiving support (non-supported papers) and up to 5 years during which support was received (supported papers) (Figure 1).

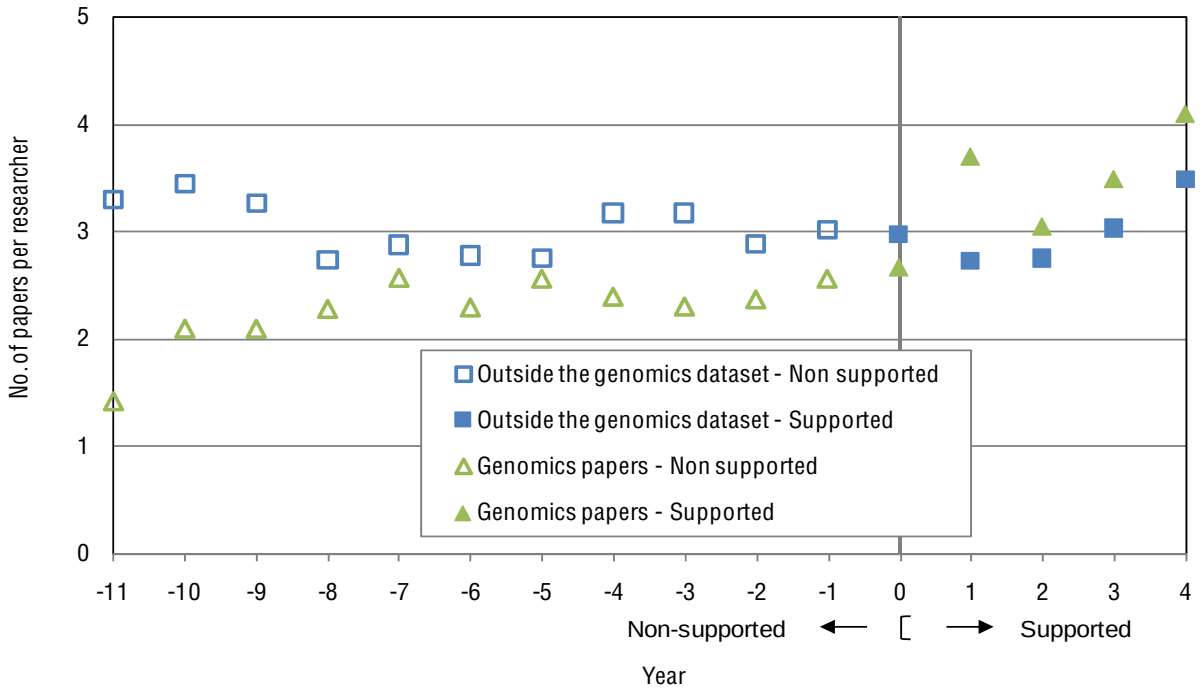

Figure 1 Trends in the number of published papers per researcher for a group of researchers who received support from Genome Canada, up to 11 years prior to receiving support and up to 5 years with support, 1996–2007

Source: Calculated by Science-Metrix using the WoS

When papers not classified within Science-Metrix' genomics dataset are considered, there is no clear trend shift between non-supported and supported papers. Nevertheless, the average number of papers produced per funded researcher is increasing within supported papers as a function of time under support by Genome Canada. When taking into account papers classified in genomics, the increase in the average number of papers produced per funded researcher appears to be stronger for supported than for non-supported papers. This pattern suggests that Genome Canada's funding has had a positive effect on the production volume of the researchers it funded. In fact, the average number of papers produced each year increased for nearly 65% of the researchers during the period when they were receiving funds from Genome Canada, and there is a significant difference between the production of the supported researchers before and while they were funded by Genome Canada (two-sample paired [Wilcoxon] signed rank test; data not shown).

However, if Genome Canada supported a number of young and promising researchers whose scientific production usually exhibits exponential growth early in their career, this pattern might have been observed when analyzing the same group of researchers without Genome Canada's support. In order to examine this possibility, it would be necessary to tag non-supported papers as being published before or after the period of support. A decrease in the production of funded researchers following the period of support would confirm Genome Canada's positive impact on the production of the researchers it funded. Unfortunately, not enough time has elapsed since Genome Canada's first competition (2001) to analyze non-supported papers after the period of support (see Section 2.2.2).

The results also provide evidence that Genome Canada's supported researchers are increasingly involved in genomics research, with the percentage of their production classified in this field (based on Science-Metrix' genomics dataset) having increased from 30% eleven years prior to receiving support from Genome Canada to 54% after five years with support (Figure 2). There is no clear shift in the trendline between non-supported and supported papers, so that concentration of output in genomics might have had increased at a similar pace without support from Genome Canada. Importantly however, there is a significantly greater proportion of genomics papers among supported papers (53%) than among non-supported papers (44%).

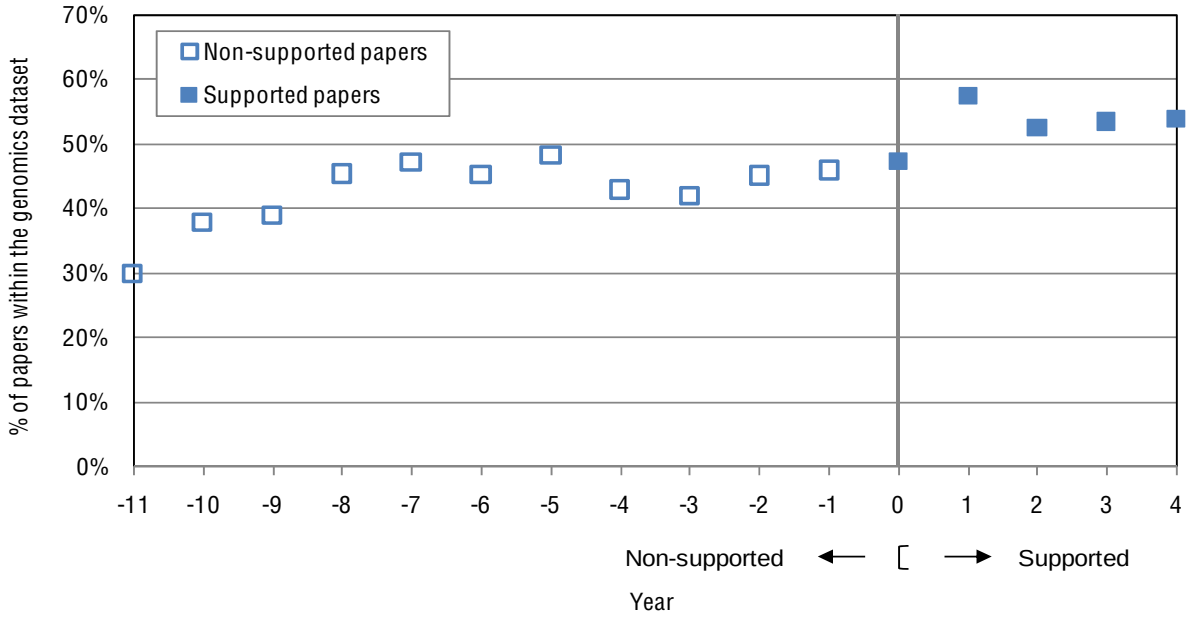

Figure 2 Trend in the percentage of genomics papers by researchers who received support from Genome Canada, up to 11 years prior to receiving support and up to 5 years with support, 1996–2007

Source: Calculated by Science-Metrix using the WoS

3.3 Scientific impact of Genome Canada's supported researchers while they were receiving funds from Genome Canada and while they were not

The observed (ARC) and expected (ARIF) scientific impact of genomics papers authored with financial support from Genome Canada are significantly higher than those of non-supported genomics papers (i.e., authored by Genome Canada's supported researchers while they were not receiving funds from the agency) (Table II).

Table II Comparison of the observed (ARC) and expected (ARIF) scientific impact of Genome Canada researchers' supported and non-supported papers in the genomics dataset, 1996–2007

Group	ARC*	ARIF
Non-supported	1.8	1.5
Supported	2.1	1.6
***p*-value**	**0.04**	**0.01**

Note: * Relative citation counts are unreliable at this analytical level for the most recent years (2006 and 2007); therefore, papers published in those years were not included in computation of the ARC indicator (see Section 2.3).

Source: Calculated by Science-Metrix using the WoS

The observed difference can be attributed to the fact that Genome Canada's supported researchers produced a greater proportion of high impact papers while they were receiving funds from Genome Canada than while they were not. This is noticeable in terms of observed impact, as there is a significantly greater proportion of supported than of non-supported papers within the upper

quartile of the citation distribution of world genomics papers (Table III). This is also noticeable in terms of expected impact, as there is a significantly greater proportion of supported than of non-supported papers in the 5% of world genomics papers with the highest impact factors (Table IV).

Table III Proportion of Genome Canada's supported and non-supported papers within quartiles of the citation distribution of world genomics papers and within the 5% most cited genomics papers in the world, 1996–2005

	No. of papers		% of papers		
Percentile interval	**Non-supported**	**Supported**	**Non-supported**	**Supported**	**Z-test for two proportions**
0-25	169	75	0.14	0.16	NS
25-50	278	82	0.23	0.17	$P < 0.01$
50-75	337	116	0.28	0.25	NS
75-100	424	198	0.35	0.42	$P < 0.01$
5% most cited	117	59	0.10	0.13	NS
0-100	1,208	471	1.00	1.00	NA

Note: Relative citation counts are unreliable at this analytical level for the most recent years (2006 and 2007); therefore, papers published in those years were not included in calculating the ARC indicator.
Source: Calculated by Science-Metrix using the WoS

Table IV Proportion of Genome Canada's supported and non-supported papers within quartiles of the distribution of world genomics papers' impact factors and within the 5% of world genomics papers with the highest IF, 1996–2007

	No. of papers		% of papers		
Percentile interval	**Non-supported**	**Supported**	**Non-supported**	**Supported**	**Z-test for two proportions**
0-25	345	133	0.16	0.13	$P < 0.05$
25-50	476	217	0.22	0.21	NS
50-75	523	301	0.24	0.29	$P < 0.05$
75-100	799	400	0.37	0.38	NS
5% highest impact factor	215	131	0.10	0.12	$P < 0.05$
0-100	2,143	1,051	1.00	1.00	NA

Source: Calculated by Science-Metrix using the WoS

However, not all researchers who received financial support from Genome Canada produced high impact papers during the period that they were funded. In fact, it is very likely that only a portion of those researchers contributed to the higher impact of the supported versus the non-supported papers, as more than half of them have experienced a decrease in their observed (58% of researchers) and expected (52% of researchers) impact during the period when they were receiving funds from Genome Canada. If one looks at the effect of Genome Canada's funding on the impact of individual researchers, instead of looking at its effect on the group of supported researchers as a whole (as in Table II), there is actually no significant difference in the impact (both observed and expected) of the supported researchers before and while they were funded by Genome Canada (two-sample paired [Wilcoxon] signed rank test; data not shown).

It would be tempting to conclude that the increase in the scientific impact of supported versus non-supported papers is the result of Genome Canada's funding. However, it is not impossible that, albeit the statistically significant difference in impact, this difference may be fortuitous. In fact, the difference in observed and expected impacts can be attributed to the high impact of a mere 1.5% of supported papers and, moreover, is attributable to a small fraction of the supported researchers. As

such, the overall impact of supported papers could actually vanish if such high impact papers were not produced again in the years to come. Considering that the scientific impact (both observed and expected) of supported papers is slowly declining, this conjecture is not improbable (Figure 3).

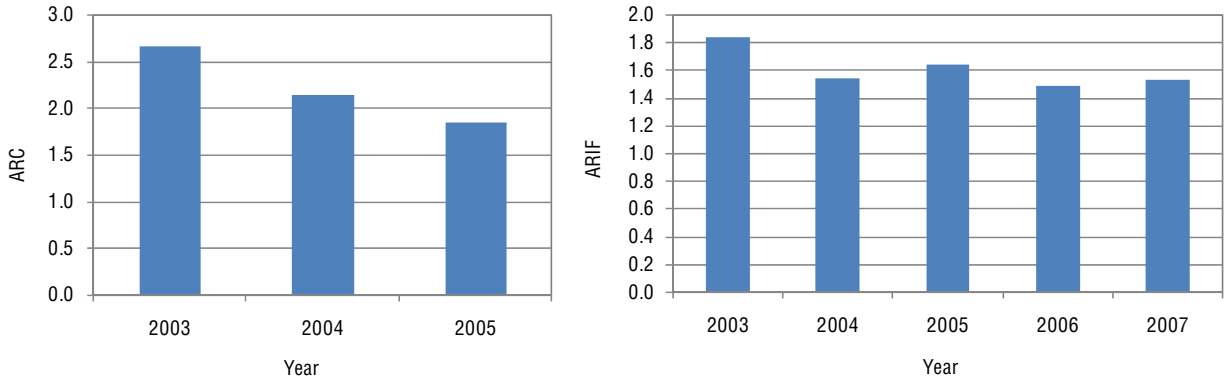

Figure 3 Observed (ARC) and expected (ARIF) scientific impact of supported papers

Note: The ARC of papers in 2006 and 2007 are not presented since they are unreliable at this analytical level.

Source: Calculated by Science-Metrix using the WoS

As additional data on supported papers and non-supported papers published after the period of support become available, the effect of Genome Canada funding on the scientific impact of the papers it has supported will become clearer. In particular, a decrease in the scientific impact of the papers authored by the supported researchers after their period of funding by Genome Canada would provide strong evidence in support of its funding positive effect.

3.4 Success stories—highly cited papers published during the period of support by Genome Canada

Table V lists the 10 most cited papers among those produced by researchers during the period while they were receiving funds from Genome Canada (for papers published between 2003 and 2005). These papers may include, in addition to articles published as a result of projects funded by Genome Canada, papers authored during the period of Genome Canada support but as a result of other projects funded by other agencies. Among these high impact papers are success stories that would not likely have been accomplished had Genome Canada not been created.

Table V The top 10 papers in terms of citation scores among the papers authored by researchers during the period while they were receiving funds from Genome Canada, 2003–2005

Source	Title	TC	RTC	Acknowledges GC	GC-funded tech. platform*
Science, 2003, 301(5633): 653-657.	Genome-wide Insertional mutagenesis of Arabidopsis thaliana	901	29.4	No	No
Nature Genetics, 2004, 36(9): 949-951.	Detection of large-scale variation in the human genome	368	26.4	Yes	Yes
Science, 2003, 300(5624): 1399-1404.	The genome sequence of the SARS-associated coronavirus	777	25.4	Yes	Yes
Nature, 2004, 428(6982): 493-521.	Genome sequence of the Brown Norway rat yields insights into mammalian evolution	522	23.3	No	Yes
Science, 2004, 303(5659): 808-813.	Global mapping of the yeast genetic interaction network	443	19.8	Yes	Yes
Nature Genetics, 2004, 36(3): 299-303.	A tiling resolution DNA microarray with complete coverage of the human genome	260	18.7	Yes	Yes
Nature Genetics, 2004, 36(5): 471-475.	Functional variants of OCTN cation transporter genes are associated with Crohn disease	217	15.6	Yes	Yes
PLoS Biology, 2003, 1(2): 166-+.	The genome sequence of Caenorhabditis briggsae: A platform for comparative genomics	104	15.0	No	Yes
PNAS, 2002, 99(26): 16899-16903.	Generation and initial analysis of more than 15,000 full-length human and mouse cDNA	606	14.9	No	Yes
Science, 2005, 309(5733): 436-442.	The genome of the kinetoplastid parasite, Leishmania major	197	14.1	No	Yes

Note: * Work was executed at a centre with a GC-funded technology platform; TC = Total citations; RTC = Relative total citations

Source: Calculated by Science-Metrix using the WoS

A prominent example of this is the large-scale project executed by the Michael Smith Genome Sciences Centre to sequence the entire genome of the SARS virus. This project would never have been executed so rapidly without the funding and infrastructure made accessible through Genome Canada. This undertaking was described by Lim in a 2003 article, published in *Symbiosis*:

> In a tour de force of genomics, government research centers in Canada and the U.S. decoded the genome of the coronavirus [...] The British Columbia Cancer Agency (BCCA) in Vancouver was the first to sequence the SARS genome [...] To sequence the SARS genome, the genome was broken in manageable fragments. Within a week, all the fragments had been sequenced [...] The Canadians took considerable pride in narrowly beating the U.S. Centers for Disease Control and Prevention (CDC) in the race to sequence SARS. [11]

In addition, the paper on the sequence of the SARS genome, published in *Science* in 2003, has already had a tremendous impact within the scientific community. It ranks among the top 500 most cited genomics papers published between 1996 and 2005. In fact, only 0.05% of genomics papers (338 out of 622,767 papers) rank above it.

Apart from the classical performance indicators (e.g., number of published papers or scientific impact [ARC, ARIF]), such success stories undoubtedly contribute to the prestige of Canada and its recognition as a world leader by its international peers. Indeed, whereas Canada was already among

[11] Lim, H.A. 2003. Bioinformatics, nanotechnology & SARS. *Symbiosis*, pp. 26-29. (http://www.dtrends.com/Nanotech/nanotech.html)

leading countries in genomics in the mid 1990s and has since maintained its position,[2] it was not yet recognized as such by other international leaders. Canada lacked an organization capable of supporting large-scale projects of high scientific merit that would attract attention and result in success stories. As was stated by international leaders interviewed within the context of the interim evaluation of Genome Canada, the creation of Genome Canada made this a possibility:

> [Genome Canada] has been successful in helping Canada catch up to most of the international leaders in genomics research. The vast majority of international peer reviewers had a strong opinion about how Canada's genomics research efforts were considered almost inexistent from an international perspective three years ago and how, today, Canada has not only "emerged as an international player" but has surpassed some of the established leaders, which "in such a short period of time", is recognized as a major achievement.[3]

3.5 Attracting outstanding scientists to Canada

The substantial financial resources made available through Genome Canada for funding large-scale genomics projects appear to have been a factor in attracting outstanding scientists to Canada. For example, Dr. Ken Dewar, co-author of the groundbreaking papers on the first draft of the human genome sequence,[2] and Dr. Thomas Hudson, co-author of one of the most highly cited genomics papers of the last decade,[12] were both at the Whitehead Institute for Biomedical Research—Center for Genome Research prior to moving to Canada with financial support from Genome Canada. Dr. Hudson went on to found one of the six genome centres—the McGill University and Genome Québec Innovation Centre—established by Genome Canada.

This outcome (attracting prominent researchers to Canada with Genome Canada support) is reflected in the bibliometric analysis of papers published by researchers who received funds from this organization to pursue large-scale genomics projects. In fact, there is a significantly greater proportion of papers with a Canadian address among the papers authored by these researchers while they were receiving funds from Genome Canada than among their papers authored while not under the support of the funding agency (Table VI). This higher proportion of papers with a Canadian address may also be the result, in part, of increased levels of collaboration between foreign scientists who have received funding from Genome Canada and Canadian researchers.

Table VI Percentage of genomics papers with a Canadian address among non-supported and supported genomics papers, 1996–2007

	No of papers		
Dataset	**With a Canadian address**	**Total**	**% of papers with a Canadian address**
Non-supported	1,816	2,143	85%
Supported	1,008	1,051	96%
Z-test for two proportions	-	-	$P < 0.01$

Source: Calculated by Science-Metrix using the WoS

[12] Daly, M.J. 2001. High-Resolution Haplotype Structure in the Human Genome. *Nature Genetics*, 29: 229-232.

Conclusion

This study aimed at providing quantitative data in support of the second evaluation of Genome Canada. More specifically, performance measurements regarding the execution of and outcomes resulting from its third objective—to support the most promising genomics projects to be performed by outstanding researchers that could not have been funded through existing mechanisms given their scope and scale—are presented and discussed.

By comparing the scientific impact of Genome Canada's supported researchers to that of the average Canadian and world researchers (excluding Genome Canada's supported researchers) in genomics, Science-Metrix established strong evidence in support of Genome Canada's successful implementation of its peer-review process for the selection of outstanding researchers. Indeed, it was found that the papers produced by researchers funded by Genome Canada had significantly higher observed and expected scientific impacts than other genomics papers from Canada and the world.

Overall, the funding of large-scale genomics projects by Genome Canada seems to have had a positive effect on the scientific performance of the group of researchers it supported. The bibliometric analysis of the scientific production of Genome Canada's supported researchers revealed that the pace of increase in the average number of genomics papers produced per researcher has intensified during the period of support compared to the preceding period. The production of genomics papers actually increased for nearly 65% of the researchers with support from Genome Canada and the difference in their production before and while they were funded by Genome Canada is statistically significant. However, care should be taken in interpreting these results, as they may be explained by factors other than Genome Canada's funding. For example, if Genome Canada supported a number of young and promising researchers whose scientific production typically exhibit exponential growth early in their career, this pattern may have been observed when analyzing the same group of researchers without Genome Canada's support.

Besides the likely beneficial effect Genome Canada has had on the scientific performance of the group of researchers it supported, the current analysis highlighted two positive outcomes resulting from the creation of Genome Canada. Firstly, among the high impact papers produced with financial support from Genome Canada are success stories that would not likely have been accomplished given their scale had the funding agency not been created. These papers disclose major scientific achievements that have had tremendous impact within the scientific community in a very short period of time, and some of them rank within the top 0.05% of the most cited genomics papers in the world. Based on the opinion of international peer reviewers interviewed within the context of the interim evaluation of Genome Canada[3] and the results of the present study, these success stories undoubtedly contributed to the prestige of Canada and its recognition as a world leader by its international peers.

Additionally, the results of this study suggest that the substantial financial resources made available through Genome Canada for funding large-scale genomics projects have been a factor in attracting outstanding scientists to Canada and might have also promoted increased levels of collaboration

between foreign scientists who have received funding from Genome Canada and Canadian researchers.

Importantly, one of the most important finding is that no significant difference was found between the scientific impact of the supported *researchers* before and while they were funded by Genome Canada. However, the overall scientific impact of *papers* authored with financial support from the funding agency is significantly higher than that of the non-supported papers. This seemingly contradictory result is explained by the fact that not all researchers with financial support from Genome Canada contributed to the higher impact of the supported versus the non-supported papers. In fact, even though the difference between the impact of the supported and non-supported papers is significant, it is modest and can be attributed to a small percentage of high impact papers. As such, the overall impact of supported papers could actually vanish if such high impact papers were not produced again in the years to come. This is not improbable considering that the scientific impact (both observed and expected) of supported papers is slowly declining.

As additional data on supported papers and non-supported papers published after the period of support become available in the coming years, it will be possible to assess more precisely the effect of Genome Canada on the scientific performance of the researchers it supported to confirm the preliminary evidence of its positive effect that was revealed by this bibliometric assessment. For example, the detection of a decrease in the production of funded researchers and/or a decrease in the scientific impact of their papers after their period of funding by Genome Canada would provide strong evidence in support of the funding agency's positive effect.

This study would not be complete if one did not note the conundrum associated with gearing funds towards excellence. The current model of many Canadian agencies is to finance the best researchers—those with a proven track record. It is clearly a challenge for the best and most established researchers to further improve both their productivity and the impact of their research. In this context, it is not surprising to find that funding the best produces great results, but it does not change the research landscape profoundly nor does it fuel the achievement of astonishingly greater scientific excellence. As such, it might be relevant to attribute a part of the funds to teams composed of promising researchers (i.e., with innovative projects) rather than to researchers who are already outstanding. This is surely more risky and selection models for these researchers are few and far between, but it might be one of the most potent ways to inject vitality in the research system.