International Agency for Research on Cancer

World Health Organization

Scientific Council **SC/53/8**
Fifty-third Session 02/12/2016

Lyon, 25–27 January 2017
Auditorium

IARC STRATEGY AND PLANS FOR BIOINFORMATICS

1. Background on bioinformatics within the IARC Medium-Term Strategy (MTS)

The field of bioinformatics is making an increasingly important contribution to cancer research. Recent technological and analytical advances allow for the unprecedented description of the molecular mechanisms involved in cancer development. Similarly, data sharing across the scientific community is creating a vast array of *in-silico* resources. Both have enormous potential in IARC's multi-disciplinary studies but also rely heavily on bioinformatics to deal with these often complex datasets. As such, bioinformatics has an important role in the inter-disciplinary research approaches outlined in IARC Medium-Term Strategy (MTS) and one that is very complementary to traditional epidemiology, biostatistics and laboratory sciences.

IARC's Bioinformatics activity predominantly stems from genomics (MCA, ICB, MPA)[1], genetics (GEN) and metabolomics (NME), providing the data to generate (or support) hypotheses regarding the molecular mechanisms of carcinogenesis and how these interplay with environmental exposures to exert their effects on the carcinogenic process. There is also potential for the early detection and monitoring of cancer, e.g. using circulating tumour DNA (ctDNA), as well as exposure biomarkers (e.g. metabolomics-based markers).

In the above context, bioinformatics involves the handling of massively-parallel sequencing and mass spectrometry data, to process, analyse, store, annotate and explore these complex datasets. Standard and established processing and analytical approaches may be applied, but there is also need for a parallel development of innovative bioinformatics and statistical techniques in support of IARC's specialized applications. Bioinformatics also touches on important routine activities of the Agency such as sample management (the SAMI platform), maintenance of scientific records (electronic notebooks, ELN) and the maintenance of the IT hardware and software needed to support these activities.

The current objective is to strengthen bioinformatics as a component within the inter-disciplinary mission outlined in the MTS. As a relatively small institute with a broad remit, IARC must ensure that the resources are used to their fullest capacity, focused on IARC's particular mission and remain complementary to resources which can be accessed through our external collaborative partners.

[1] See Annex 1 below for list of acronyms

2. Investment and strategy to-date (current status)

2.1. Unified oversight structure – Bioinformatics Steering Committee (BISC)

In recognition of its increasing importance at IARC, in 2013 the Director requested advice on current and projected bioinformatics activities. An internal group was convened to evaluate the status of bioinformatics at the Agency and identify areas that could be improved.

The internal review favoured developing Bioinformatics as a devolved, matrix-style model as opposed to a dedicated bioinformatics service group. This model saw bioinformaticians being nested within the scientific groups. This allows for the necessary specialization while building upon the existing multidisciplinary staffing structures in place within IARC's scientific groups. The internal review also advised on the computing investment needed.

A decision was also made to provide a level of Agency-wide coordination of bioinformatics, to create links between bioinformaticians and thereby extending their potential impact beyond the group level, representing a second dimension in the "matrix". An overarching structure, the **Bioinformatics Steering Committee (BISC)**, was therefore created to oversee bioinformatics at the Agency. BISC is mandated with the ongoing development of IARC's bioinformatics capability and to advise the Director regarding developments in this area (see BISC Terms of Reference and membership list)[2]. BISC comprises and receives advice from two active and dynamic working groups (WG), the **Bioinformatics Working Group** and the **Informatics Working Group**.

The Bioinformatics WG's role is to facilitate interactions between bioinformaticians at multiple levels: by undertaking regular meetings aimed at sharing bioinformatics-related skills and knowledge, and coordinating training activities for bioinformaticians, particularly for post-doctoral fellows/students. This WG is led by IARC P-level scientific staff and made up of a much more diverse group of scientists, post-doctoral fellows and students working in the area of "omics"-related analysis.

The Informatics WG oversees the informatics hardware / software and monitors requirements in support of the IARC bioinformatics activities, and provides advice to the BISC (and in turn to the Director) on future developments and investments. Its membership is made up of the technical staff related to the IT infrastructure at the Agency.

2.2. Hardware

Between 2011 and 2016, IARC has expanded in a strategic manner its scientific IT capability to address the projected demand for bioinformatics-related computing and to avoid saturation of the existing IT systems. Key contributions were made through the Governing Council Special Fund in 2011 and again in 2013 (Document GC/55/14F), with additional funds obtained through IARC's Groups (scientific and ITS) and through extrabudgetary sources. Examples include:

A medium-scale high performance computing cluster (HPC). This system was put in place in 2011 and has gone through two sequential upgrades that have been carried out over the last five years.

[2] See Annex 2 below for BISC ToR and Annex 3 for BISC membership

These allowed expanded massively-parallel sequencing analyses performed by a number of Groups/Sections.

On-demand private cloud computing using an "Infrastructure as a Service" (IaaS) structure. The IaaS model is a form of private cloud computing that provides virtualized computing resources using OpenStack technology. It consists of a central infrastructure hosting shared hardware, software and storage resources. Using a web-based interface, IARC scientists can then generate a virtual machine provisioned with resources appropriate to the task (in terms of CPU, memory and storage). IaaS is well-suited for IARC workloads that are often temporary, experimental or that change unexpectedly. This system was initially supported by a communal regular budget contribution across eight IARC groups and additionally upgraded with Governing Council Special Funds.

Upgrades to IARC's IT infrastructure. These were undertaken by ITS using IARC regular budget funds. Two upgrades were performed over the last five years. The LAN Network has been completely replaced and internet access upgraded to appropriately access *in-silico* databases (TCGA, UK Biobank).

2.3. Personnel

The increasing importance of bioinformatics expertise for the Agency has necessitated some staffing changes in scientific groups. This has included reorientation of scientific positions: a P3 and LY5 in GCS (M. Foll and C. Voegele); promotion of current staff (to P3 for M. Olivier in MMB, duties include oversight of TP53 database); and creation of research assistant positions (LY5 in EGE/MMB). IARC has also added P2 staff members familiar with bioinformatics to GCS, GEP (supported with external funds) and BMA (2013–2015). There are also several biostatistics and database management staff now working on aspects that relate to bioinformatics (LY6 GEP, LY5 ENV), and an LY4 database manager (sample management SAMI) has also been created in LSB.

IARC's non-permanent staff (post-doctoral fellows and students) make a very important contribution to bioinformatics. This includes a variety of early career scientists that use bioinformatics within their research projects, but also four dedicated early career scientists that are working directly in bioinformatics. Despite the increased investment in staffing in recent years, there is a recognition within the Agency that additional investment is needed, an opinion supported by the recent Scientific Council Review Panel evaluations of the MCA and GEN Sections.

2.4. Outside collaborations

A number of important collaborative links have been formed in support of bioinformatics projects. The most prominent example is the collaboration with the *Plateforme de Bioinformatique Gilles Thomas* of the Synergie Lyon Cancer foundation (SLC, Centre Léon Bérard, Lyon). SLC (headed by Professor Gilles Thomas, until his sad passing in 2013, and now led by Alain Viari) provided IARC with additional HPC capacity and these interactions greatly facilitated the development of particular massively-parallel sequencing protocols established at IARC. The SLC staff members

are also active attendees of the IARC BISC "Omics" seminar series, and maintain active collaborations between SLC and IARC.

Other examples of important, project-centred external ties are IARC's active collaborations in the fields of chromatin structure dynamics (MMB, with Department of Biomedical Informatics, Ohio State University, OH), cancer genomics (GEP, ICGC CAGE-Kid), epigenome research (EGE with Medical Epigenomics Laboratory at the CeMM Research Center for Molecular Medicine, Austrian Academy of Sciences) and mutational signatures of environmental carcinogens (MMB, GCS) with the Duke-NUS Centre for Computational Biology, Singapore and the Theoretical Biology and Biophysics laboratory at the Los Alamos National Laboratory, NM. In addition, the CLINGEN initiative of the US NCBI has built a working group on the classification of TP53 variants to assess pathogenicity, in which the IARC TP53 database participates as collaborator.

The Agency adopts an approach of in-house expertise complemented by research collaborations on specific subject areas as well as the more strategic collaborations within Lyon.

3. Achievements to-date

3.1. Analytical workflows

Bioinformatic analyses generally consist of a prescribed series of processing steps called a pipeline or a workflow. The increasingly complex laboratory techniques available (like next-generation sequencing or mass-spectrometry) have intensified the need for robust and reproducible pipelines as they become more sophisticated. Implementing and deploying either standard established pipelines, or the ones developed internally in support of IARC's specialized applications is not a trivial task and there is a need for dedicated frameworks to do so. In the past two years we have implemented several workflows using mainly two platforms: "nextflow" scripts for users willing to use a command line interface and a "Galaxy" server workbench suited for researchers with limited informatics skills needing a graphical interface.

These platforms support regularly used bioinformatics pipelines for analyses of the DNA methylome, genome-wide mutational signatures, NGS alignment, germline and somatic variant calling. All the pipelines developed are shared across IARC groups and more widely with the community through the IARC Github organization page with currently seven contributors (https://github.com/IARCbioinfo) as open-source repositories. Particular attention is put on ease of use, reproducibility and scalability, allowing analyses to be performed easily on personal computers, HPC clusters or in the cloud. This aspect is especially important to ensure that scientists from low- and middle-income countries can also benefit from IARC development in bioinformatics.

3.2. Scientific outputs

Bioinformatics at IARC is now contributing to a wide variety of IARC's projects conducted across different Groups/Sections, resulting in high-quality papers, as exemplified by recent published studies on etiology and signatures of environmental factors (Scelo et al., Nat. Commun. 2014; Castells et al., CEBP 2015; Ambatipudi et al., Epigenomics 2016), genetic susceptibility (Wang et al., Nat. Genetic 2014; Lesseur et al., Nat. Genetic 2016) mechanisms and mechanistic models (Olivier et al., Sci. Rep. 2014), early detection (Fernandez-Cuesta et al., EBioMedicine 2016; Le Calvez-Kelm et al., Oncotarget 2016), TP53 database (Bouaoun et al., Hum Mutat 2016), design of bioinformatics tools (ELN, (Voegele et al., Bioinformatics 2013), MutSpec ,(Ardin M et al., BMC Bioinformatics 2016), Needlestack (Fernandez-Cuesta et al., EBioMedicine 2016), and metabolomics studies (Zamore-Ros et al., Sci. Rep. 2016)). Pioneering chemo-informatics approaches supporting the IARC's carcinogen evaluation and classification programme have also been developed (Guha et al., Environ. Health Perspect. 2016).

3.3. Training

The specialized needs of bioinformatics at the Agency suggest that organic growth of existing staff in bioinformatics may be beneficial, particularly as many current staff have existing skills relevant to bioinformatics (for example skills in statistical software such as R). Specific training initiatives could be offered to personnel (for example training in the R statistical genomics package "Bioconductor"). The member groups of the BISC have coordinated several courses in bioinformatics, working in collaboration with ETR, "Basic UNIX for handling large datasets" (25 March 2013, Participants: 29, June 15–17, 2016 Participants: 25), "Basic Training sessions on R" (7, 14 & 21 October 2015. Participants: 17), "BioConductor for Integrative Genomic Analyses" (26–28 October 2015. Participants: 24) and "Introduction to high performance computing (HPC) and the IARC linux cluster" (June 17, 2016. Participants: 25). Training sessions on the Galaxy bioinformatics tools are planned in early 2017.

3.4. 'Omics' seminar series

The Bioinformatics WG of the BISC has been organizing bi-weekly seminars with the participation of IARC and Centre Léon Bérard bioinformaticians and a broader audience. This has resulted in discussion-based presentations generally attended by between 25–30 participants. The aim of these seminars is to promote scientific exchanges on 'omics' applications (from next generation sequencing to metabolomics), bioinformatics methods, to share work-in-progress results and perspectives in informal presentations and discussions.

4. Future requirements

Bioinformatics is continuing to increase in importance at IARC. There are now particular areas in which the Agency needs to reinforce, noted by Section peer reviews, the Scientific and Governing Councils and in internal IARC discussions. Hardware, storage and high-performance computing demand continues to grow and strengthening is needed to ensure that the scientific needs are met. This includes laboratory methods, but also the infrastructure related to IT. In addition to

extensive consultation across research groups within the Agency the Director also convened and Advisory Group on bioinformatics to inform future investment plans; the Group comprised two Scientific Council members and two additional external experts[3].

Following consultation the key areas in which IARC plans to expand its capacity in bioinformatics are described below.

4.1. Hardware: computing processing, storage

With the increase in importance of bioinformatics, there has been a rapid increase in needs for high performance computing (HPC) and storage at IARC. The current computing capacity mostly comes from the HPC cluster purchased in 2012 initially with 96 computing cores, expanded in 2015 with 192 additional cores and further updated in 2015 with a dedicated storage server.

The Agency progressed from four users of this system in 2014 to 19 active users in September 2016 from six different scientific groups (GCS, GEP, MMB, ENV, ICB, and NMB). In the past six months, around 500 000 hours of computing have been performed on the cluster, corresponding to an average use of 50% of the full capacity, and the volume of data being stored and backed-up has increased accordingly. This level of usage is now approaching the maximum capacity which is realistic in practice, as there is a need to support peaks and avoid long waiting times for analyses.

The Agency therefore envisages the need for a doubling in the HPC and storage capability over the next 18 months. Request for funding for this additional equipment (endorsed by the Advisory panel[3]) will be presented to the Scientific Council in January 2017 (see document SC/53/6) **and to the Governing Council for consideration in May 2017.** In addition, the Agency is continually using collaborative options for analytical capacity whenever possible. Moving forward, the IT WG of the BISC continues to consider the cost benefits of internal capacity and moving towards a hybrid model including cloud-based solutions to cover utilization peaks.

4.2. Massively-parallel sequencing capacity

In respect to specialized laboratory equipment, namely in the “omics” domains, there is a need to maximize efficiency in-house through common platforms and facilities that are typically created and managed centrally as shared resources. Direct access to such instrumentation enables recruitment and retention of leading molecular scientists.

IARC scientists currently use three small-throughput DNA sequencers, allowing a maximum sequencing output ranging from 2 Gb to 15 Gb. These sequencers are well-suited for low-complexity sequencing strategies but their capacity does not allow for more complex applications

[3] Composition of the Advisory Panel: *Scientific Council (SC) members* – Stephen Chanock & Lukas Huber; *External Experts* – Roland Eils (DKFZ, Heidelberg) & Ivo Gut (Genome Analysis National Center, Barcelona) *IARC* – Christopher P. Wild, Director; Matthieu Foll (Scientist (Bioinformatics), GCS); Zdenko Herceg (Head, Section of Mechanisms of Carcinogenesis (MCA) and Head, Epigenetics Group (EGE)); Christopher Jack (Informatics Officer, Information Technology Services (ITS)); James McKay (Head, Genetic Cancer Susceptibility Group (GCS));’ Augustin Scalbert (Head, Biomarkers Group (BMA)); Jiri Zavadil (Head, Molecular Mechanisms and Biomarkers Group (MMB))

where cost-effectiveness can be achieved by higher-level multiplexing. Upon a broad discussion in-house **the Agency has identified the need for a versatile desktop sequencer with a capacity of 100–120 Gb** (see document SC/53/6**). A request for funding for this additional equipment will be made to the Governing Council in May 2017, depending on evaluation by the Scientific Council in January 2017.** The additional equipment would allow the cost-effective implementation of the novel sequencing applications needed by multiple research groups at IARC. This equipment would be operated as shared resource.

At the same time, IARC does not intend to harbour higher-throughput sequencing instrumentation in-house and will instead seek strategic external partnerships for such high-capacity needs. Access to one or two preferred suppliers will be explored in order to negotiate external contracts for high-throughput sequencing with the best service portfolio and pricing.

4.3. Personnel

A devolved, matrix-style staffing model within IARC's Groups/Sections remains suited to the Agency, and it is to be maintained and expanded. Key positions in several scientific groups that require additional support have been identified. The specialized nature of each of the scientific groups justifies such nested scientific and technical positions, and a close collaboration and coordination is envisaged between these individuals to allow the matrix structure to harness potential synergies.

In parallel to scientific support, IT systems administration should be reinforced at the technical level, with the benefit of freeing time from professional scientific staff currently implicated in such support tasks. The matrix model will be upheld for the next two to three years and relevant achievements in this period will be closely monitored and assessed. Possible alternative models, as discussed by the Advisory Group[3], such as a research-mandated Computational Biology Group, will remain open for further consideration; the potential creation and added value to IARC's mission may be re-examined in line with regular restructuring at IARC.

Particular areas to be supported in the next phase are as follows:

- a computational Professional grade biologist within BMA for analysis of metabolomics data;
- additional technical (LY-grade) staff in two scientific Sections/Groups (MMB/EGE and ICB), to assist the existing scientific staff with high-end analysis of genomic data;
- a dedicated IT technician support (LY-grade) for the hardware infrastructure and software engineering.

These changes will ensure that the bioinformatics expertise – ranging from technical to scientific levels – will be distributed across the Sections/Groups in a balanced manner, to support their day-to-day needs as well as long-term goals. The funding for these additional positions has come from abolition of existing regular budget posts in other areas, with reallocation of resources to bioinformatics, as well as utilization of funds currently available to the Director from the Governing Council Special Fund.

5. Strengthening of the cross-Agency activities

The expansion of the devolved, matrix model for bioinformatics also requires a strengthening of cross-Agency exchanges in bioinformatics, beyond the current activities of the BISC and its WGs. Collaboration is critical to the successful roll-out and sustained, effective operation of the matrix model. Collaborative activities thus should be integral to all IARC's bioinformaticians, and this aspect will be included as defined tasks within these positions.

Much of the cross-Agency coordination will continue to be overseen by the BISC. However, in order to provide additional leadership, the P3 GCS Bioinformatician will be tasked with a dedicated coordination role, with responsibility to promote collaboration across the Agency. This function will encourage interactions between the bioinformaticians to boost knowledge, creativity, technique-benchmarking and expertise-sharing as well as harnessing synergies across the Agency's activities wherever possible. The role is also designed to ensure that newly developed methodologies are made available across IARC and to the scientific community in general. The new tasks are enabled through shifting duties regarding IT management to the IT positions described above.

5.1. Multi-disciplinary collaboration

It is important to recognize that bioinformatics at the Agency works within a multidisciplinary structure and should thus not be considered in isolation. Bioinformatics can draw upon IARC's existing expertise in genetics, epidemiology, laboratory skills, information technology and biostatistics. Many aspects of these fields overlap with the broad nature of tasks associated with the word bioinformatics. As such, all interactions between these fields are to be fostered.

An example of how IARC is to achieve these interactions is the revision of the structure of the BISC. Before the end of 2016, the Agency will create an autonomous yet complementary Biostatistics Working Group positioned within the BISC structure. This will reinforce the links between bioinformatics and biostatistics at IARC. This will alter somewhat the role of the Steering Committee by shifting it closer to a higher-level oversight body, with each of the WGs undertaking their activities in a more autonomous manner. One of the important roles of the revised BISC will be to facilitate and maintain the collaborative links with strategic external partners.

5.2. External collaboration

That IARC maintains capacity and direct access to advanced technologies and bioinformatics is important for fostering scientific activities, and to maintain the Agency's scientific competitiveness and attractiveness to high-quality staff. Nonetheless, IARC has maintained a policy of strategic partnerships with local centres of expertise, in order to avoid redundancy and overspecialization. Balancing and enhancing IARC's collaborations with centres with greater bioinformatics capacity is equally important and more formal links could be established with groups such as the *Plateforme de Bioinformatique Gilles Thomas* of the Synergie Lyon Cancer foundation at the Centre Léon Bérard, Lyon. Collaborative and capacity-sharing models with Synergie Lyon Cancer may also provide access to highly trained bioinformatics staff. Within France, one such initiative includes the France Médecine Génomique project, a large medical sequencing operation to be

built over the next five years. IARC is in preliminary discussions with local Lyon-based partners regarding potential collaborations in this context.

6. Expected achievements

While appropriate resourcing of bioinformatics as a highly dynamic field poses challenges in fully capturing the research potential of the technological advances, making strategic decisions to strengthen the field of bioinformatics is important for the success of IARC's scientific mission and its interdisciplinary activities. The above-described investments, reorganization and developments will substantially augment the capacity for research in the key areas to be pursued at the Agency over the next three years, providing unprecedented, detailed insights into cancer etiology, streamlining prevention opportunities as well as improving avenues to early detection of cancer.

Recognizing the constant evolution of opportunities and requirements in this field, the Agency anticipates consulting further with the Scientific Council and external experts on this topic in the future.

ANNEX 1

Sections and Groups (as at 1 July 2016)

Acronym	Full name of Section/Group	Responsible Officers
CSU	**Section of CANCER SURVEILLANCE**	**Dr F. Bray**
EDP	**Section of EARLY DETECTION AND PREVENTION**	**Dr R. Herrero**
PRI	Prevention and Implementation Group	Dr R. Herrero
SCR	Screening Group	Dr Sankaranarayanan
ENV	**Section of ENVIRONMENT AND RADIATION**	**Dr J. Schüz** Deputy: Dr A. Kesminiene
GEN	**Section of GENETICS**	**Dr P. Brennan**
GCS	Genetic Cancer Susceptibility Group	Dr J. McKay
GEP	Genetic Epidemiology Group	Dr P. Brennan
IMO	**Section of IARC MONOGRAPHS**	**Dr K. Straif** Deputy: Dr D. Loomis
INF	**Section of INFECTIONS**	**Dr M. Tommasino**
ICB	Infections and Cancer Biology Group	Dr M. Tommasino
ICE	Infections and Cancer Epidemiology Group	Dr S. Franceschi
MCA	**Section of MECHANISMS OF CARCINOGENESIS**	**Dr Z. Herceg**
EGE	Epigenetics Group	Dr Z. Herceg
MMB	Molecular Mechanisms and Biomarkers Group	Dr J. Zavadil
MPA	**Section of MOLECULAR PATHOLOGY**	**Dr H. Ohgaki**
NME	**Section of NUTRITION AND METABOLISM**	**Dr M. Gunter**
BMA	Biomarkers Group	Dr A. Scalbert
NEP	Nutritional Epidemiology Group	Dr M. Gunter
NMB	Nutritional Methodology and Biostatistics Group	Dr P. Ferrari

ANNEX 2 – BISC Terms of reference

International Agency for Research on Cancer

World Health Organization

IARC Bioinformatics Steering Committee (BISC)

Terms of Reference

The role of the Committee will be as follows:

I. Consider how the broad field of bioinformatics can be used to support and advance the scientific activities of the IARC.

II. To provide strategic and technical advice to the Director regarding the development of IARC's bioinformatics related activities (in keeping with the IARC Medium-Term Strategy).

III. To oversee, and receive advice from, two interconnected working groups:

- **The Bioinformatics Working Group**, whose responsibility will be to facilitate interaction, to promote skill development and knowledge sharing in bioinformatics within the Agency and relevant collaborative partners.

- **The Informatics Working Group**, whose responsibility will be to study the information technology aspects of the Agencies scientific activities; consider IT solutions aimed at the scientific activities, monitor for IARC's informatic infrastructure (hard- and soft- ware) for scientific activities and anticipate potential future requirements.

IV. To provide advice and recommendations to DIR regarding training opportunities for IARC personnel within the area of bioinformatics.

V. Facilitate the development of optimal bioinformatic resources by promoting the interaction of informatics, statistics, laboratory sciences and epidemiology.

ANNEX 3 – BISC List

IARC - MEMORANDUM

From: Dr C.P. Wild, DIR **To:** All Staff **Date:** 27/01/2014

Ref.: **Subject:** Creation of the IARC Bioinformatics Steering Committee

---

I am pleased to announce the creation of the IARC Bioinformatics Steering Committee (BISC) which will oversee progress in capacity in bioinformatics at IARC and ensure that the Agency responds to new developments in a timely and informed manner.

Dr James McKay will act as Chair of the BISC. The other committee members, who will serve for an initial term of two years, are as follows:

Dr Jiri Zavadil, MMB (Vice-Chair)
Dr Graham Byrnes, BST
Mr Christopher Jack, ITS
Dr Mattias Johannson, GEP
Dr Dinesh Kumar, BMA
Dr Hiroko Ohgaki, MPA
Dr Magali Olivier, MMB
Dr Martyn Plummer, ICE
Dr Massimo Tommasino, ICB
Dr Hector Vargas, EGE
P3 (TBA), GCS