International Agency for Research on Cancer

World Health Organization

Scientific Council **SC/53/6**
Fifty-third Session 25/11/2016

Lyon, 25–27 January 2017
Auditorium

PURCHASE OF SCIENTIFIC EQUIPMENT

1. Laboratory sciences, biostatistics, bioinformatics and epidemiology are closely integrated in the IARC Medium-Term Strategy. This requires high quality laboratories and the availability of state-of-the-art scientific equipment. Constant upgrade and acquisition of scientific instruments are thus essential to support this strategy.

2. Centralized platforms (e.g. next-generation sequencing (NGS) or mass spectrometry) are available to scientists across the Agency and used to generate genomic, transcriptomic, epigenomic or metabolomic data. A high success rate in obtaining competitive research grants over the last few years has resulted in a significant increase in the number of samples analysed on these platforms, and in the volume of data acquired and analysed.

3. A new plan has been made to support developments in bioinformatics (see document SC/53/8) needed to interpret the complex datasets generated. Acquisition of some new equipment is also needed to support these rapidly growing research activities.

4. The Director would like, therefore, to request the Governing Council at its 59th session in May 2017, to provide an allocation of €700 000 from the Governing Council Special Fund to purchase the following equipment:

a) An upgrade of the IARC scientific computing capacity.

b) An upgrade of the IARC next-generation sequencing (NGS) platform.

c) The acquisition of an automated system to study cancer chromatin at genome-wide level.

5. This approach is first submitted to the Scientific Council for its consideration.

6. The annual maintenance costs of the requested equipment will be covered by the regular budget as well as by collaborative programmes through grant applications.

a) Upgrade of the IARC scientific computing capacity

7. 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.

8. The current computing capacity at IARC mostly comes from the High Performance Computing (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 of 60 TB and an additional 100 TB dedicated to archives and backups. IARC scientists also have access to a private cloud environment that provides on-demand virtualized computing resources using OpenStack technology. It consists of a central infrastructure, hosting shared hardware (32 cores, 512 GB of RAM), software and storage resources, dedicated to general purpose computation and development.

9. The Agency progressed from four users of the above system in 2014 to 19 active users in September 2016, from six different scientific Groups (GCS, GEP, MMB, ENV, ICB, and NMB)[1]. In the six months prior to preparing the current document, around 500 000 hours of computing have been performed on the cluster and the volume of data being stored and backed-up has increased accordingly. The current level of usage is approaching the maximum capacity, recognizing the need to provide for peaks in demand and to avoid long waiting times for analyses.

10. In order to address these pressing needs there is a requirement to establish a modular scalable cluster with a minimum of double the current CPU cores, RAM and storage capacity. The cluster will be scalable by adding computing or storage nodes to the environment to increase performance or capacity. In addition to this, a dedicated backup and archive environment will be established to protect important raw data and analysis results.

11. Maintenance of 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, for example with the "*Plateforme de Bioinformatique Gilles Thomas"* of the Synergie Lyon Cancer foundation at the Centre Léon Bérard, Lyon. In addition, the Agency uses collaborative options for bioinformatics capacity whenever possible. The IT Working Group of the Bioinformatics and Biostatistics Steering Committee (BBISC) continues to advise the Director on the cost benefits of internal computing capacity and adopting a hybrid model including cloud-based solutions to cover utilization peaks.

12. The proposed equipment would be operated as a shared resource under the responsibility of the IT Working Group of the BBISC, who would provide access and support for other research groups at IARC.

13. Estimated cost: €300 000 (allocated as follows: computing servers: €140 000; storage: €50 000; backup: €30 000; general infrastructure (network, power supply, cables, etc.): €80 000).

[1] GCS = Genetic Cancer Susceptibility Group;
GEP = Genetic Epidemiology Group;
MMB = Molecular Mechanisms and Biomarkers Group;
ENV = Section of Environment and Radiation;
ICB = Infections and Cancer Biology Group;
NMB = Nutritional Methodology and Biostatistics Group.

b) Upgrade of the IARC next-generation sequencing (NGS) platform

Benchtop sequencer of 100–120 Gb capacity

14. An important line of research conducted by several groups at IARC involves the successful identification and functional evaluation of genetic and epigenetic alterations in tumour cells, pre-neoplastic lesions or normal tissues. For this research, IARC scientists currently use small-throughput sequencers, allowing a maximum sequencing output of 2 Gb data for 200–400 nt single reads (Ion PGM), up to 15 Gb for 200 nt single reads (Ion Proton) or up to 15 Gb for 2 x 300 bp paired-end reads (MiSeq). These sequencers are well-suited for low-complexity sequencing strategies but their capacity does not allow for more complex applications where cost-effectiveness can be achieved by higher-level multiplexing.

15. Consequently, a desktop sequencer with a capacity of 100–120 Gb would allow the cost-effective implementation of novel sequencing applications now required by multiple research groups. The sequencer will enable rapid access for IARC research groups to a number of sequencing strategies of critical interest, which are currently only available through rather fragmented third-party arrangements. A cost comparison (including staff time, consumables and maintenance) of the proposed equipment compared to an outsourcing option demonstrated the better value of the current proposed investment. In contrast, high-throughput applications (e.g. whole genome and large exome studies) will continue to be performed through strategic (external) partnerships.

16. The proposed 100 Gb capacity sequencer will notably enable research on:

- Mutational signature analyses in support of mechanistic carcinogen evaluation (in vitro mutagenicity assays, bioassay tumours, mainly using low coverage exome sequencing (MMB, IMO, GCS, GEP).[2]
- Multiplexed ChIP-seq and chromosome structure and conformation analysis (EGE, MMB, ICB).
- RNA sequencing (MMB, EGE, ICB, GCS).
- Reduced representation bisulfite sequencing (EGE).
- Structural alteration mapping, including analysis of fusion genes (GCS).
- Viral integrations and 16S metagenomics (ICB).
- Discovery of non-invasive and specific biomarkers based on nucleic acids in body fluids (GCS, GEP, MMB).
- Cancer driver gene mutation screening by targeted resequencing in highly multiplexed sample sets (GCS, MMB, EGE, MPA).

17. The proposed equipment would be operated as a shared resource overseen by existing laboratory technicians under the responsibility of EGE, MMB and GCS, who would provide access and support for other research groups at IARC. Dedicated Information Technology Services (ITS) support is available to this equipment. Data analysis and storage will be accommodated with the above request for an expansion of ITS infrastructure and computational capacity.

18. Estimated cost: €310 000

[2] IMO = Section of IARC Monographs;
EGE = Epigenetics Group;
MPA = Section of Molecular Pathology.

c) Automated system to study cancer chromatin at genome-wide level

19. Additional support is needed to maintain and upgrade IARC's capacity to perform analyses of chromatin at the genome-wide level, as the current number of robotics-based chromatin studies is expected to increase four- to five-fold over the coming years. Notably the scale of studies is increasing as the approaches find more routine applications within large-scale projects, thus requiring specialized robotics to automatically and efficiently perform sample preparation, and to provide high-quality data at a reduced labour cost and greater reproducibility and efficiency. Specific areas of activity include:

- Chromatin immunoprecipitation combined with NGS (ChIP-seq) (EGE, MMB, ICB).
- Protein binding, protein-protein and protein-DNA interactions (EGE, MMB, ICB).
- Other applications include: MeDIP, hMeDIP, MethylCap, Re-ChIP, MagBisulfite, RNA-IP, IPure.

20. One automated robotics system is available at IARC (Diagenode SX-8G IP-Star). However, the workload increase outlined above will soon exceed the instrument's capacity. The proposed system will increase the existing capacity and simultaneously allow for efficient and cost-effective library preparation for NGS platforms. This is an important consideration given the proposed upgrade of NGS and would be of interest to multiple research groups (EGE, MMB, GCS, ICB). At the same time it meets the increased need for the cost-effective in-house preparation of libraries, irrespective of a sequencing upgrade.

21. Estimated cost: €90 000

Requested budget

	Approximate price (€)
a) Upgrade of the IARC scientific computing capacity	300 000
b) Upgrade of the IARC next-generation sequencing (NGS) platform	310 000
c) Automated system to study cancer chromatin at genome-wide level	90 000
Total equipment	700 000