ARMENIAN BIOINFORMATICS INSTITUTE

SCIENTIFIC EDUCATIONAL FOUNDATION

FUNDRAISER

FOR THE ESTABLISHMENT OF A

GENOMICS & BIOINFORMATICS

COMPUTING INFRASTRUCTURE

FOR

ARMENIA

EXECUTIVE SUMMARY

Bioinformatics is a scientific discipline that encompasses all aspects of biological information acquisition, processing, storage, distribution, analysis, and interpretation. As data has become central to biology research, the need for bioinformaticians – scientists who handle these data – is becoming more and more critical. Naturally, activities setting up standards and capacity for biological big data collection, distribution and analysis are cornerstones for strategic developments and for shaping the future of biomedicine and biotechnology.

Conducting bioinformatics research requires extensive computing infrastructure: both in terms of hardware (large storage and computing capacity, fast read-write operations, large operational memory, and parallelization), as well as in terms of cloud access and software support. So far ABI has joined forces with IMB and IIAP institutes in Armenia to provide computing resources not only to researchers (>12) but also to students (>30). These resources, however, have reached their maximum usage capacity.

More and more students and researchers in various universities and institutes in Armenia, including ABI, require access to bioinformatics computing infrastructure. This underlies our aim to create unprecedented opportunities for bioinformaticians by setting up a computing infrastructure that will be used by students and scientists alike.

The proposed computing infrastructure will serve research and educational purposes and will support around 20 researchers and more than 40 students initially. For its implementation, we will obtain a hardware set-up with 128 cores, 48 TB hard drive storage, and 1TB of random access memory (with an opportunity for further expansion), as well as hire and train a system administrator to support cloud sharing and software/database setup specific for bioinformatics applications, in collaboration with our international partners.

BIOINFORMATICS: WHY IS IT ESSENTIAL FOR ARMENIA?

Medicine, pharma, and biotech have largely been transformed into data-dependent ventures, where large-scale molecular, clinical, and population data are cornerstones of success. Technologies that have revolutionized bio-related research and development, such as large-scale sequencing, have posed the challenge of having enough bioinformaticians able to transform data into knowledge. Recent surveys report that only 50% of demand in bioinformaticians is fulfilled in academia and industry worldwide. This gap between the availability and the need for a workforce continues to grow.

The first group in Armenia specializing in genomics & bioinformatics (Bioinformatics Group, Institute of Molecular Biology NAS RA), established ten years ago, has shown a successful track record of conducting original research and acquiring valuable collaborations and attracting significant funding from national and international agencies. However, the speed of organic growth at BIG wasn't sufficient to produce enough bioinformatics manpower. Since its establishment in 2011, BIG has produced several alumni (now postdoc, PhD, and MSc elsewhere), and currently has two PhD students. Today, the amount of genome bioinformatics specialists in the country that hold a PhD degree is only two.

As a consequence, many labs conducting research in life sciences and medicine in Armenia are in extreme need of bioinformatics specialists. Despite the presence of educational programs with partial coverage of bioinformatics subjects, those only prepare seeds for bioinformatics researchers. Today, the number of people with a PhD degree in genome bioinformatics in the country is just 2, instead of at least 30 that are needed in the academic, medical, and industrial sectors according to international standards. Even more troubling is the fact that presently the educational programs in Armenia do not prepare genomics/bioinformatics specialists at a solid level and in sufficient numbers to satisfy the needs that academic, medical, and industrial entities will have in the near future. There is a gap between the content of educational programs and the latest advancements in bioinformatics, as well as a lack of academic entities supporting the transition of students to master's, PhD, and postdoctoral levels.

The only way to overcome this problem is to initiate a combined educational/research program based on available local and international expertise to build a minimum fundament of the required human capital.

With this aim, in February 2021, a team of bioinformatics specialists from Armenia and from abroad established the Armenian Bioinformatics Institute (ABI) as a non-profit scientific educational foundation. ABI plans to develop the human capital in bioinformatics required to boost data-driven research and innovation in life sciences including biomedicine and biotechnology. ABI serves as a platform to unite experts from around the world, and to recruit students to the exciting field of bioinformatics, connect them to qualified mentors and supervisors, and provide a fruitful environment for learning, research, and networking.

Contact: lilit.nersisyan@abi.am *Armenian Bioinformatics Institute*

February 21, 2022

STAKEHOLDERS AND PROJECTS

Several research and educational institutions in Armenia need bioinformatics training and support, and most of them need temporary or permanent access to computing infrastructure to perform their research or computations needed for their courses or training activities. Below we summarize estimates for 2022.

- Armenian Bioinformatics Institute (a growing user base with currently 12 researchers, estimated to reach 20 by 2023)
- Institute of Molecular Biology National Academy of Sciences (NAS) of Armenia (Armenian atlas of cancer genomes, vine genome project, genomic surveillance of food-borne infections)
- Yerevan State University (iBOL reference genome mapping of Armenian Biodiversity, several projects in microbiome research)
- Institute of Botany NAS Armenia (mycology and lichenology research)
- Armenian National Agrarian University (selection of abiotic stress resistance traits in plants)
- Santé Arménie biological resource center
- Russian-Armenian University (research on induced pluripotent stem cells, ~ 15 students in bioinformatics per year)
- American University of Armenia (~ 5-10 students in bioinformatics per year)
- Yerevan State University (~ 2-5 students per year)

The number of organizations and the size of target groups are expected to increase year by year. With the proposed setup we will support around half of the needs described here. However, it is designed with further extension in mind and additional funds will make it possible to easily double or triple the computational capacity, as the size of human capital grows.

COMPUTING RESOURCES IN ARMENIA

Currently, ABI collaborates with the Institute for Informatics and Automation problems NAS RA (IIAP) in Armenia that provides virtual machines to the Institute of Molecular Biology (IMB) and the Armenian Bioinformatics Institute (ABI). ABI sets up bioinformatics software packages and databases on this server and provides access to up to 30 students/researchers in Armenia. However, the resources that can be allocated by IIAP (64 CPU nodes, 256 GB RAM, 10 Tb storage) have already reached their cap. To keep up with the growing needs of ABI, as well as research and educational systems in the country, it is required that we significantly expand the computing infrastructure to scale up the service for a wider community.

The government expects a supercomputer to arrive from Toulouse, however the exact date of its arrival is not known yet, and it will take a couple of years until it becomes functional and provides distributed computing access. We will apply for access on this server as well. Notably, by the experiences of our international partners, centralized “supercomputer solutions” are suboptimal for many applications. Therefore, it is urgently needed to set up a local bioinformatics computing infrastructure in Armenia independent of future supercomputer access options.

but in the meantime, obtaining a local server and training experts to set up and support a bioinformatics computing infrastructure is the best temporary solution.

The cloud computing platforms, such the Amazon Web Services, cost a few hundred thousand USD per year for the specifications mentioned below, which is nearly twice as much, as we need for a local setup (and only a one-time cost). Moreover, security requirements for human genome data of national impact exclude web services as a working option in many situations.

PROPOSED HARDWARE AND SOFTWARE ARCHITECTURE

We are planning to set up an infrastructure with 128 cores (256 threads), 1 TB of random access memory (can be extended to 2TB), 48 TB of hard drive storage (can be extended to 96 TB), and 3xTB SSD. The server will be fed by a constant power supply (see below). The usage is planned for at least 10 years, and the capacity can easily be extended upon request in the next years.

The server will be set up and maintained by a dedicated server administrator employed at ABI.

HARDWARE

SERVER INFRASTRUCTURE

We are planning to acquire a Dell PowerEdge R7525 model, with a total price of 48,450 $. Please, find below the detailed specification of the server:

Qty	Item Description	SKU
0	**Dell PowerEdge R7525 Server**	**210-AUVQ**
1	SAS/SATA Backplane	379-BDSS
1	Trusted Platform Module 2.0 V3	461-AAIG
1	12X 3.5 + Rear 2X 2.5 SAS/SATA with XGMI	321-BFDX
1	AMD 7713 2.0GHz,64C/128T,256M,225W,3200	338-BZRM
1	AMD 7713 2.0GHz,64C/128T,256M,225W,3200	338-BZRO
1	Heatsink for 2 CPU configuration (CPU greater than or equal to 180W)	412-AATC
1	Performance Optimized	370-AAIP
16	64GB RDIMM, 3200MT/s, Dual Rank	370-AEVP
1	RAID 5	780-BCDP
1	PERC H745 Controller Adapter Low Profile	405-AAWR
6	8TB 7.2K RPM SATA 6Gbps 512e 3.5in Hot-plug Hard Drive	400-ASIF
3	960GB SSD SATA Read Intensive 6Gbps 512 2.5in Flex Bay AG Drive, 1 DWPD,	400-AXRL
1	Power Saving BIOS Settings	384-BBBH
1	UEFI BIOS Boot Mode with GPT Partition	800-BBDM
1	High Performance Fan x6	750-ADGL
1	Dual, Hot-Plug,Power Supply Redundant (1+1), 1400W, Mixed Mode	450-AJHG
2	C13 to C14, PDU Style, 10 AMP, 6.5 Feet (2m), Power Cord	450-AADY

1	Riser Config 8, Half Length, 4 x16 slots	330-BBPK
1	PowerEdge R7525 Motherboard, with 2 x 1Gb Onboard LOM,MLK V2	384-BCWO
1	iDRAC9,Enterprise 15G	385-BBOT
1	Broadcom 5720 Quad Port 1GbE BASE-T Adapter, OCP NIC 3.0	540-BCOB
1	PowerEdge 2U Standard Bezel	325-BCHU
1	No Quick Sync	350-BBYX
1	iDRAC,Legacy Password	379-BCSG
1	ReadyRails Sliding Rails	770-BBBQ
1	Cable Management Arm, 2U	770-BDRQ
1	Fan Foam, HDD 2U	750-ACOM
1	3Yr Warranty	709-BBIY

CONSTANT ELECTRICITY SUPPLY

To keep the server constantly up and running, we will obtain an up to 6kWt electricity generator, working on gasoline, with ABP/ATS and AVR functions, an onLine, manageable UPS, up to 2.5 kW output, as well as a rack cabinet 24U (including cooling and power systems).

SOFTWARE ARCHITECTURE

A list of free bioinformatics software (see Appendix 1) will be installed for everyone to load and use, at the same time enabling local installation of custom software for each user. In addition, databases commonly used by the bioinformatics community (*e.g.* human and mouse genomes, microbiome classification databases, *etc*) will be built and constantly updated.

The SLURM resource distribution system will be implemented for user-submitted jobs. The user groups from various universities and research institutions in Armenia will be given project-based access and space allocation.

The system administrator will have an opportunity to get training in the HPC centers in Germany and Sweden and will have consultation support from the University of Leipzig running a similar infrastructure. In turn, the administrator will be responsible for preparing user manuals, as well as providing short training to the users about how to use the computational resources efficiently.

FUNDING NEEDED

Item	Cost, USD
Hardware infrastructure	48500
Electricity supply stabilization	4500
Support costs for 1 year (including travel)	15000

Total funds needed	68000
Funds available	10000
Funds to be raised	58000

FUTURE DEVELOPMENTS

The proposed infrastructure will serve two main purposes: (i) satisfy the current needs for computing for researchers and students, and (ii) develop the expertise and human resource capacity to scale up for larger resources in the middle and long run.

ABI has recently applied for a Horizon Europe Twinning grant call, together with the Interdisciplinary Centre for Bioinformatics at Leipzig University, Germany, and the National Bioinformatics Infrastructure Sweden (NBI.am project). One of the purposes of NBI.am is to build a national computing infrastructure for Armenia that provides computing resources, bioinformatics software, and service support to researchers and students, and that hosts a bioinformatics core facility, providing more customized bioinformatics support to academic and industrial entities. NBI.am envisions an IT staff of up to 2 people, to get training not only in Armenia but also at the supercomputing centers in Sweden and in Germany. However, the project does not envision the hardware support to ABI, which we are planning to solve via the donation route.

APPENDIX 1

LIST OF COMMONLY USED BIOINFORMATICS SOFTWARE PACKAGES

Below you may find some commonly installed software packages. This is not a comprehensive list and will be constantly updated upon request.

Software	Description
Compilers and build tools	
Git	A free and open-source distributed version control system designed to handle everything from small to very large projects with speed and efficiency.
Python 3 (does not conflict with python module)	A programming language that lets you work quickly and integrate systems more effectively.
Python	A programming language that lets you work quickly and integrate systems more effectively.
Julia	Designed from the beginning for high performance. Julia programs compile to efficient native code for multiple platforms via LLVM.
htop	A cross-platform interactive process viewer. It is a text-mode application (for console or X terminals) and requires ncurses.
swapspace	dynamic swap manager for Linux
Misc	
GNU Emacs	An interpreter for Emacs Lisp, a dialect of the Lisp programming language with extensions to support text editing.
CMake	An open-source, cross-platform family of tools designed to build, test, and package software.
VIM	A highly configurable text editor built to make creating and changing any kind of text very efficient.
Statistics	
R	A free software environment for statistical computing and graphics.
Libraries	
bzip2	a command-line file compression program
libcurl	a free and easy-to-use client-side URL transfer library
zlib	unobtrusive compression library
BLAST	An algorithm and program for comparing primary biological sequence information
BWA	A software package for mapping DNA sequences against a large reference genome

ClustalW	a series of widely used computer programs used in bioinformatics for multiple sequence alignment.
STAR	Spliced Transcripts Alignment to a Reference
HISAT2	a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes as well as to a single reference genome.
Bioinformatics Misc	
Bamtools	BamTools provides both a programmer's API and an end-user's toolkit for handling BAM files.
BCFtools	a program for variant calling and manipulating files in the Variant Call Format (VCF) and its binary counterpart BCF.
BioPerl	an international association of users & developers of open source Perl tools for bioinformatics, genomics, and life science
BioPython	a set of freely available tools for biological computation written in Python by an international team of developers.
Bismark	A tool to map bisulfite converted sequence reads and determine cytosine methylation states
gvcftools	a set of utilities to create and analyze Genome VCF (gVCF) files.
HTSlib	an implementation of a unified C library for accessing common file formats, such as SAM, CRAM, and VCF, used for high-throughput sequencing data, and is the core library used by samtools and bcftools.
Integrative Genomics Viewer	a high-performance, easy-to-use, interactive tool for the visual exploration of genomic data.
Integrative Genomics Viewer Tools	a high-performance, easy-to-use, interactive tool for the visual exploration of genomic data.
Kraken2	a taxonomic sequence classifier that assigns taxonomic labels to short DNA reads.
ngsplot	collects a large database of functional elements for many genomes.
Picard	a set of command-line tools for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.
PLINK2	a whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.
PySam	a python module for reading, manipulating, and writing genomic data sets.
RSEM	a software package for estimating gene and isoform expression levels from RNA-Seq data.
SAMtools	a suite of programs for interacting with high-throughput sequencing data
SRA Toolkit	a collection of tools and libraries for using data in the INSDC Sequence Read Archives.
UMI-tools	contains tools for dealing with Unique Molecular Identifiers (UMIs)/Random Molecular Tags (RMTs) and single cell RNA-Seq cell barcodes.
SnpEff	annotates and predicts the effects of genetic variants on genes and proteins
String Graph Assembler (SGA)	a de novo genome assembler based on the concept of string graphs.
Spades	an assembly toolkit containing various assembly pipelines
cd-hit	a widely used program for clustering biological sequences to reduce

	sequence redundancy and improve the performance of other sequence analyses.
Centrifuge	a very rapid and memory-efficient system for the classification of DNA sequences from microbial samples, with better sensitivity than and comparable accuracy to other leading systems.
Nucdiff	locates and categorizes differences between two closely related nucleotide sequences.
Bioinformatics Phylogeny	
MEGA	software suite for analyzing DNA and protein sequence data from species and populations.
Bioinformatics Pipelines	
Bowtie2	an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.
Cufflinks	assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples
cutadapt	finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads.
Tophat	a fast splice junction mapper for RNA-Seq reads.
Minimap2	a versatile sequence alignment program that aligns DNA or mRNA sequences against a large reference database.
medaka	a tool to create consensus sequences and variant calls from nanopore sequencing data.
kallisto	a program for quantifying abundances of transcripts from bulk and single-cell RNA-Seq data, or more generally of target sequences using high-throughput sequencing reads.
Nanopolish	Software package for signal-level analysis of Oxford Nanopore sequencing data.
artic-ncov environment	nCoV-2019 Nanopore sequencing protocol package
Bioinformatics SW Collections	
ensembl-vep	predicts the functional effects of genomic variants.
ANNOVAR	an efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes
BBMap	a suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data.
bcl2fastq	demultiplexes sequencing data and converts base call (BCL) files into FASTQ files.
BEDTools	a swiss-army knife of tools for a wide-range of genomics analysis tasks.
GATK	the industry standard for identifying SNPs and indels in germline DNA and RNAseq data
MEME Suite	Perform motif discovery on DNA, RNA, protein, or custom alphabet datasets.
VCFtools	a program package designed for working with VCF files