White paper

The Applied & Translational Genomics Cloud (ATGC)

Authors: Benedikt Brors[‖], Werner Eberhardt[‡], Roland Eils[‖ ¶], Jürgen Eils[‖], Nina Habermann*, Sergei Iakhnin*, Ajay Kumar[‖],Jan Korbel*, Christian Lawerenz[‖], Peter Lichter[‖], Rupert Lück*, Fruzsina Molnar-Gabor[# ¶], Tobias Rausch*, Kai Sachs[‡], Matthias Schlesner[‖], Christof von Kalle[§], Sebastian Waszak*, Joachim Weischenfeldt*

Affiliations: [‖]German Cancer Research Center (DKFZ), Heidelberg, Germany;
[‡]SAP SE, Germany;
*European Molecular Biology Laboratory (EMBL), Heidelberg, Germany;
[¶]Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany;
[#] Heidelberg Academy of Sciences and Humanities, Heidelberg, Germany;
[§]National Center for Tumor Diseases (NCT), Heidelberg, Germany

Please note the information on liability and use included in footnote '§' on page 6.

1. Introduction & Vision – the Applied & Translational Genomics Cloud (ATGC)

a. Current challenges: state-of-the-art of genomics

Recent developments in DNA sequencing technology now enable human whole-genome sequencing (WGS) for less than €1,000. These cost reductions are spurring initiatives such as the UK-based 100,000 Genomes Project, the International Cancer Genome Consortium (ICGC) with its recently initiated Pan-Cancer Analysis of Whole Genomes Project (PCAWG) and the German Consortium for Translational Cancer Research (DKTK) to pursue analyses of vast numbers of patient genomes. We estimate that >25,000 genomes will be completed by 2018 in Germany alone. Together with other molecular data types (*e.g.*, transcriptomes, microbiomes) an unprecedentedly rich set of data will emerge, facilitating integrative analyses to enable improved patient stratification, diagnostics and personalized medicine. While relevant to many diseases, cancer research is considered to have a pioneering role in this regard, given high rates of patient participation and several successes demonstrating that cancer genomic data could be used to improve patient management[1-3].

While opportunities abound, there are significant challenges: No single German university or research center currently has the necessary infrastructure to perform analyses with such large datasets, and to store and access these data securely. Further, lack of standardization in computational analysis workflows renders data processed in different institutions non-comparable. Vast infrastructure investments will be necessary to enable integrative analyses with these data and efficient utilization in research and translation.

b. Brief overview: the ATGC model

To overcome these challenges we propose the development of the Applied & Translational Genomics Cloud (ATGC), with a model of cloud computing[4] involving pooled resource utilization through responsible sharing of IT infrastructure[*] and pre-defined services to facilitate engaging non-experts. ATGC will serve as a cloud for high-throughput data in the life sciences in Germany, with initial focus on cancer genomics data, and plans for a later expansion into other data types and life science areas. ATGC's core infrastructure will adhere to current data protection standards, and will strive for technical compatibility to allow expansion within international contexts. Installation of ATGC's infrastructure will have several key advantages:

- Making genomic analysis with state-of-the-art tools widely accessible, since ATGC will provide bioinformatics processing capabilities to numerous users in Germany, including to researchers and clinicians in institutions lacking infrastructure or expertise for patient genome analysis, diagnostic laboratories, as well as biotech and pharmaceutical industries.
- Availability of pre-configured pipelines to facilitate state-of-the-art genomic analyses (Software-as-a-service, SaaS, model).
- Standardization of analysis approaches to improve comparability of datasets across institutions, to enable integrative analyses and meta-analyses.
- Improved data protection through standardized data access control and centralized data storage.
- Reduction in overall infrastructure and operational costs through resource sharing (by avoiding duplication of infrastructure).

Implementation of our ATGC model may also have positive effects on economic development. Cloud computing is about to become a 100 billion USD market[5-7], and strong commitments in Germany into cloud computing will ensure international competitiveness, improve job security, and form incentives for biotech industries. The development of ATGC would follow recommendations by the Leopoldina, cautioning that Germany can only remain competitive by strategically setting up a national "omics" and IT infrastructure linking universities with non-university institutions, to bundle expertise in interdisciplinary research[8].

[*] The principle of cloud computing is that multiple users employ a shared pool of computers and data storage devices that are accessed remotely. Clouds can be open to the public (public cloud), or confined to a defined community (community cloud). Clouds typically combine several characteristics: pooling of resources for expedited computation and elasticity (*i.e.* scalability of computes), on-demand self-service, broad network access and standaridized data protection. Service models include provisioning of Software-as-a-Service (SaaS) and Infrastructure-as-a-Service (IaaS).

2. Central conditions of the ATGC

a. Data types

ATGC will initially focus on the processing of human genomics data with cancer as the initial research emphasis. Data will include DNA sequencing, transcriptome, methylome, proteome and human microbiome data (augmented, whenever possible, with clinical data), which according to German law need to be protected to high data privacy and security standards. We envision that ATGC will be gradually extended to cover other data types, diseases and organisms. For example, we expect cardiovascular diseases to be a focus for future use, and envision that expansions can cover all research areas covered in German health centers.

b. Users & use cases

Envisioned users of ATGC will include bioinformaticians, clinicians as well as omics- and data-scientists. We foresee two principal use cases for ATGC, which will operate Software as a Service (SaaS) and Infrastructure-as-a-service (IaaS) cloud models:

- Bioinformaticians, computer scientists and software engineers develop software and content for ATGC: These users will develop new computational workflows (or improve existing workflows) to facilitate data analysis and interpretation using the ATGC. Analysts will also be able to employ custom analysis workflows to process genetic data stored within the cloud, using ATGC's IaaS model.
- Life scientists, clinicians, diagnostic laboratories, and members of biotech/pharmaceutical industries make use of ATGC's SaaS model to use pre-configured analysis pipelines, *e.g.* to identify mutations or to integrate mutation with clinical data in an interactive fashion. Data analysis may be accompanied by the upload of new data, and will benefit from the use of data previously available at ATGC as background information. ATGC will offer different means of data sharing, including a model where data are shared with all users, and a model providing a more enclosed user space that can be entered by defined community of users only.

Of note, ATGC's value will significantly extend beyond these two principal use cases: We foresee that basic researchers or clinicians will consult ATGC's data portal to investigate preprocessed data, for example, to obtain information on how often late-stage cancers exhibit mutations in certain genes, and how often these genes are linked with a certain metagenomic state. ATGC will further enable collaborative research by facilitating control cohort and meta-analyses, for example, use of pre-processed ATGC raw genomic data as a control cohort for cardiovascular disease genetics.

Results from these analyses can lead to the awareness that, in the absence of this knowledge, the patient concerned would be subject to significant personal harm. In this case it can be imperative that the users intervene. The ATGC enables communicating such knowledge pertaining to a specific person to the physician treating that patient, provided that the patient's statement of consent does not state otherwise.

c. Stakeholders and decision-makers

Creators and operators of ATGC will take the lead to establish a genomics cloud service in Germany. This should build on existing strong competence both in sequencing and in operating large-scale IT infrastructure for the life sciences with respect to storage as well as to compute environments, and should factor in expertise from commercial industries that are already provisioning computational clouds in the life sciences. The main responsibilities for operating ATGC are by nature organizational, and their competences comprise designing and operating ATGC, which in this function also includes conducting cloud governance. Clinicians and biomedical researchers who upload, analyze and share data through ATGC, and also bioinformaticians who develop software, shall qualify as ATGC′s users. Users will have restricted responsibilities referred to their specific activities. ATGC operators will define on a case-by-case basis who is authorized or responsible for which activity and qualifies, *e.g.* as data controller or processor.

Funders will have a vital interest to allow for access to large-scale datasets and associated compute resources, based on scientific excellence alone. If no cloud infrastructure exists, access will only be possible at those institutions that can afford to provide and maintain such infrastructures themselves. On the other hand, investments into community cloud infrastructure will make it necessary to develop models where

usage of the cloud can be paid for as a fee-for-service using research funding (rather than granting sums for IT investments, whose long-term maintenance and operation may be much more difficult to achieve).

d. Data security requirements & compliance with data protection regulation

Stakeholders and decision makers will preserve the rights (in particular the right to withdraw consent) of sample donors, whose contribution will be indispensable. Notably, ATGC's data security plan, and especially the protection of data in the cloud, will comply with stringent German data protection regulations and standards[†]. Furthermore, to allow for secure use of compute infrastructure, ATGC's data security plan software will encompass protection techniques against threats (*e.g.*, "hackers") according to the state-of-the-art in science and technology. We recommend that proven cloud certificates, and certificates covering compliance with data protection regulations, will be acquired in this context. Additional awareness is needed considering that the European Data Protection Directive 95/46/EC will be superseded by the General Data Protection Regulation prospectively by 2016. Resulting from its legal quality, this regulation will have immediate effect on all EU member states after the two-year transition period will have expired. A data access committee shall take decisions on data access applications for datasets residing within ATGC's infrastructure.

Over and above legal compliance the authors of this white paper believe that an investigation of upcoming normative issues, beyond those already studied in related contexts[9-11], shall also be decisive under ethical aspects to achieve sound solutions for shaping and conducting ATGC. We hence recommend normative research in this context to be pursued continuously.

3. Implementation of the ATGC

a. Governance and code-of-conduct

We envision that a new non-profit association[‡] not bound to a single existing research institution will be founded, with the objectives to establish and operate ATGC. This organization will also be responsible for compliance with data security and data protection requirements, which will require careful and up-to-date governance. To enable operation of ATGC in accordance with law, a code of conduct will be specified and implemented under labor laws, which will guide cloud operators as well as all users, cooperation, and industrial partners and regulate questions in reasonable detail, including on the sharing and control of sensitive data. An accountability approach provided by ATGC's code of conduct will not only provide compliance [with data protection regulations] on a practical level, but also provide trust within the research community and data donors.

b. Technical & software development requirements

To establish ATGC, investments in infrastructure, cloud services, and content development will be required.

With regard to infrastructure, driving factor for the design will be the data set size. Each human genome occupies between 150-300 Gigabytes (GB) of space on disk, and to enable analysis of paired tumor/normal genomes from 25,000 cancer patients ~15 Petabytes (PB) of disk storage will be required, a number that does not yet include mirroring of the data to comply with industry standards (to achieve the latter, 40-50 Petabytes will be necessary). Data processing capabilities amounting to 20,000 compute cores will ensure feasibility of the cloud services provisioned by ATGC, which include Software-as-a-service (SaaS) and Infrastructure-as-a-Service (IaaS) models. ATGC's data warehouse will need to support several operations associated with data governance, including mechanisms for data deposition, retrieval, encryption, access control, and provenance tracking. To enable efficient data processing, rapid connectivity in the order of tens of Gbit/s will be required between the underlying storage and the computational nodes. The whole infrastructure setup needs to be designed in a scalable fashion, to facilitate future expansions.

With regard to cloud services, access to ATGC should occur through a secure web portal. A fine-grained user management system is required to grant users access to ATGC's resources in agreement with the stakeholders' requirements. An access control layer at file system level should be used to enable

[†] Due to international differences in the stringency of data protection, especially relating to clouds, we regard it is in the best interest of German institutions to have ATGC initially implemented as a national solution in Germany, to allow continued use of research data from German patients for research and translation.

[‡] Foundation of a new non-profit association, similar to how the DFN-Verein ("Deutsches Forschungsnetz"; see https://www.dfn.de) is organized, will facilitate set up, operation and governance of ATGC.

management of data access regulations for each dataset. The web portal should also provide functions for collaborative research, for exchange of results, and for exchange of analysis workflows. The web portal will be connected to a workflow management system, thereby allowing the execution of preconfigured content such as analysis pipelines by users without computational background.

With regard to content development, ATGC will provide pipelines and preconfigured analysis workflows (SaaS model). Workflows employed by SaaS users will be developed by bioinformaticians and computer scientists from the German scientific community, and shared through ATGC's web portal or included in the workflow management system. For workflow developers, ATGC will offer an opportunity to reach a broad German user community. Usage metrics for each workflow will be provided, metrics that can be regarded as key indicators of performance and may be recognized by funders and publishers in the future, creating further incentives for workflow development and improvement. Data analysis results produced by SaaS workflows can be made available to other users, to reduce duplication of analysis of the same datasets.

c. Industry partners

For operation of ATGC, hardware, software, and facility management will be necessary. IT infrastructure could be provided from industry partners with experience in cloud computing technology, which are willing to comply with the governance by ATGC's to be founded non-profit association, and with the code of conduct. Comprehensive experience in computational genomics will be of pivotal importance, and partnerships with leading cloud service and IT providers highly beneficial to meet technological challenges. We additionally envision partnerships with biotech and pharmaceutical industries, with patient genomic data being already actively analyzed in biotech and pharmaceutical research industries. We foresee mutual benefits: *e.g.* IT providers will be able promote software and technology, whereas biotech/pharmaceutical research industries would receive prioritized access to large biomedically relevant research data.

d. Required investments and pay-on-demand models

Initial substantial investments as seed funding are needed to establish the primary infrastructure of ATGC, which comprise storage, computing, and network resources. These resources will be purchased and/or leased. Funding for these investments will be provided through (*i*) German federal funding agencies, (*ii*) foundation funds from public or private institutions, (*iii*) and/or private sponsorship. Realization of the investments primarily through these funding sources will cover the startup costs and hence enable an affordable payment mode or business model for the users. Sustainability of ATGC will require building upon a stable business model, we envision that sustainability will require repeated infrastructure reinvestments to ensure that ATGC remains competitive and effective in the future.

We foresee three payment modes, as pay-on-demand models serving interests of different users groups:

i) Researchers developing tools or workflows will pay for the use of infrastructure based on data processing time required (infrastructure-as-a-service model, IaaS).
ii) Researchers using existing tools or workflows will pay for the processing of data with workflows made available through ATGC (software-as-a-service model, SaaS).
iii) Researchers pay for usage of information and data sets released within ATGC.

Depending on the available seed funding, SaaS and IaaS services may initially be made available free of charge for users from German academic/research organizations. We envision that in the future, grants for research work on genomics could be partially re-invested in cloud infrastructure instead of sourcing individual infrastructure. To form incentives for participation, users/organizations could earn ATGC credits when submitting new datasets to ATGC. Likewise credits could be provided for the provision of technological innovation, *i.e.*, improved pipeline technologies. We expect that customers from biotech/pharmaceutical industries are willing to pay for the access of preprocessed datasets, the use of preconfigured analysis workflows (SaaS model), and the ability to process ATGC's datasets with their own customized analysis pipelines, which can be implemented using ATGC's IT infrastructure (on the basis of ATGC's IaaS model). ATGC will also allow commercial partners/SMEs to establish their own cloud service using ATGC's infrastructure, whereby these cloud services would significantly benefit from data and software resources available through ATGC.

e. Future expansions of service portfolios and users

The operators of ATGC will provide frequent training opportunities as well as workshops, to increase community participation and make newcomers from academic institutions (or industry) familiar with ATGC.

Once ATGC has been established as a credible, reliable and sustainable resource in Germany, we see great potential for new types of users and expanded service portfolios. ATGC will be set up in such a way that ensures compatibility with international research. In the mid- to long-term, the user base of ATGC may be extended to other European countries, in particular smaller countries that lack a comparable IT infrastructure, but perform relevant research studies which translational research in Germany, mediated through ATGC, may benefit from. In the future, we could further imagine diagnostic use cases relying on ATGC world-class services. Health care insurance companies could pay part of the diagnostic performance in the long term. These future extensions shall always consider the developed basic model and the underlying standards.[§]

References

1. Iyer G, Hanrahan AJ, Milowsky MI, et al. Genome sequencing identifies a basis for everolimus sensitivity. *Science* 2012; **338**(6104): 221.
2. Li T, Kung HJ, Mack PC, Gandara DR. Genotyping and genomic profiling of non-small-cell lung cancer: implications for current and future therapies. *J Clin Oncol* 2013; **31**(8): 1039-49.
3. Wu YM, Su F, Kalyana-Sundaram S, et al. Identification of targetable FGFR gene fusions in diverse cancers. *Cancer Discov* 2013; **3**(6): 636-47.
4. Well P, Grance T. The NIST Definition of Cloud Computing. *National Institute of Standard and Commerce*; **800-145**.
5. http://www.bvp.com/blog/bvp-cloud-computing-index-crosses-100-billion-market-milestone.
6. http://www.forbes.com/sites/oracle/2013/01/17/cloud-industry-rockets-toward-100-billion-mckinsey-bullish/.
7. http://www.news-sap.com/idcs-top-10-global-trends/.
8. Leopoldina. Report on Tomorrow's Sciences. Life sciences in transition - challenges of omics technologies for Germany's infrastructures in research and teaching. Halle (Saale) Germany: Deutsche Akademie der Naturforscher Leopoldina e.V.; 2015.
9. Dove ES, Joly Y, Tasse AM, et al. Genomic cloud computing: legal and ethical points to consider. *Eur J Hum Genet* 2015; **23**(10): 1271-8.
10. http://www.uni-heidelberg.de/md/totalsequenzierung/informationen/mk_eurat_position_paper.pdf.
11. http://genomicsandhealth.org/.

§ Notes on liability and use

The authors of this scientific document, having formulated it on behalf of the [named] institutions, conduct their scientific work on the basis of their usual due diligence and based on the state of science and technology known to them at the time of the work being conducted.

This document constitutes a solution model for data processing in genome research. Its informative value is naturally limited to aspects and questions which can be described in the context of the conditions on site. This document does not serve as a basis for contracts which are to be concluded at a later date. In particular, the usability of the statements which are made is to be newly negotiated.

It must also be noted that the understanding of data protection under European law is still in a considerable state of flux in many areas. In this regard, the proper application of European data protection regulations remains a dynamic task which requires attention to and, if necessary, reaction to future legal developments which could not be taken into account during the development time frame.

As part of the functions assigned to them, the authors carry out tasks in the field of applied research and venture into new areas of technology. The risks associated with this entail, in particular, the possibility of research goals not being achieved or being only partially achieved. Regardless of any envisaged subsequent conclusion of contracts, no usability of the solution model is guaranteed. It is also noted that any use of the solution model may only occur in excerpts and only under the condition of prior consultation with the involved institutions. In the event of use, the provided solution model does not release the user from the requirement to conduct their own independent legal review.