



Project Final Report

Grant Agreement number: **262055**
Project acronym: **ESGI**
Project title: **European Sequencing and Genotyping Infrastructure**
Funding Scheme: **Combination of CP & CSA**
Date of latest version of Annex I against which the assessment will be made: **21.07.2014**
Periodic report: **1st 2nd 3rd**
Period covered: **from 01.02.2011 to 31.07.2015**

Name, title and organisation of the scientific representative of the project's coordinator:

Dr. Sascha Sauer – Max Planck Gesellschaft zur Foerderung der Wissenschaften e.V.

Tel: **+49 30 8413 1661**
Fax: **+49 30 8413 1960**
E-mail: **sauer@molgen.mpg.de**
Project website address: **www.esgi-infrastructure.eu**

Table of Content

Section 1 – Final publishable summary report.....3

1.1 Executive summary 4

1.2 Summary description of project context and objectives 5

1.3 Description of the main S&T results/foregrounds of ESGI 7

1.4 The potential impact 26

Section 1 – Final publishable summary report

ESGI



Project title: **European Sequencing and Genotyping Infrastructure**

Website: <http://esgi-infrastructure.eu/>

Contractors involved (ESGI consortium):

The project is coordinated by Dr. Sascha Sauer [Partner 01] MPG Max Planck Gesellschaft zur Foerderung der Wissenschaften e.V., Germany, Address of the Coordinator, Ihnestrasse 63-73, D-14195 Berlin, Germany.

Other partners and team leaders:

Partner 02	CAU	Christian-Albrechts-Universitaet zu Kiel, Germany
Partner 03	WTSI	Genome Research Limited, Great Britain
Partner 04	EMBL-EBI	European Molecular Biology Laboratory, Germany
Partner 05	CEA-CNG	Commissariat à l'énergie atomique et aux energies alternatives, France
Partner 06	INSERM	Institut National de la Santé et de la Recherche Médicale (INSERM), France
Partner 07	PCB	Fundació Privada Parc Científic de Barcelona, Spain
Partner 08	CRG	Fundació Privada Centre de Regulació Genòmica, Spain
Partner 09	UU	Uppsala Universitet, Sweden
Partner 10	MUG	Medizinische Universitaet Graz, Austria
Partner 11	GABO:mi	GABO:mi Gesellschaft für Ablauforganisation :milliarium mbH & Co KG, Germany

1.1 Executive summary

The European Sequencing and Genotyping Infrastructure (ESGI) project was a successful joint effort of leading European sequencing, genotyping and bioinformatics centres to defragment the European research capability in a key area of the life sciences. ESGI provided external users with top-level access facilities to pursue demanding large-scale DNA analyses using first-class technologies for a broad range of genetics and systems biology applications.

In particular, massively-parallel sequencing technologies were essential components to unravel basic gene-regulating molecular pathways and disease processes, as well as to gain a better understanding of the evolution of European populations and disease susceptibility by deciphering the impact of variation of our genomes on clinical phenotypes.

The summarised accomplished goals of ESGI comprised development of an integrated distributed infrastructure of 7 major genome research sites and 4 complementing partner sites that:

- Provided state-of-the-art sequencing and genotyping systems including bioinformatics support for excellent genetics and systems biology research in Europe, which resulted in 29 transnational access projects of external users leading to a number of high-impact studies.
- Provided an infrastructure to routinely sequence the complete human genome in a few hours for less than 1000 Euro.
- Provided a flexible sustainable entity with a diverse application range in genetics and systems biology.
- Provided support and training in genome research for scientists in the 27 EU member states to contribute to a modern, science- and technology-based economy and society.

1.2 Summary description of project context and objectives

Background and Aims

Genetics and genome research are central disciplines in the life sciences and in biomedical development. High-throughput sequencing and genotyping technology platforms are important components to advance this competitive field of research.

Notably, large-scale genomics facilities (as were made accessible by ESGI) were previously not sufficiently available in many EU states. As the European Strategy Forum for Research Infrastructures (ESFRI) pointed out in the past, the development of a high-performance infrastructure for genome research was of crucial importance to position Europe as one of the world-leading regions for genetics, genomics and systems biology research. A biology research infrastructure based on sequencing and genotyping was thus considered as a key contribution to the European Research Area.

Since substantial investments in scientific and technological equipment are needed to build high-performance infrastructures for genome research, a main objective of ESGI was the integration and further development of world class high-throughput sequencing and genotyping facilities to defragment European activities in this field of research. ESGI in particular aimed to provide support to external users to generate sequence data rapidly and to acquire knowledge efficiently, using the most powerful methods available. Thereby ESGI aimed to strengthen the European research capacities in genetics and genomics and to improve the knowledge transfer from large genomics centres among themselves and to external expert groups or scientists focusing on specific research questions.

Work strategy and general description

ESGI's activities consisted of three main components, namely networking activities, R&D activities, and most importantly provision of transnational access to external users.

Networking activities supported the integration and standardisation of the various technological and methodological approaches including emerging methods and data storage formats. Agreed standards were applied for spreading good practices, and of course for well-controlled application of transnational access projects.

Networking activities were additionally implemented to rapidly transfer technology developments derived from RTD activities of ESGI partners. Moreover, a number of online-tools and training activities were offered to educate (potential) users for the demanding analysis of large-scale sequence data.

The networking (and management) activities were further designed to strengthen existing links to complementary European infrastructure initiatives, such as infrastructures for biobanking of biological samples (e.g. BBMRI) and for managing biological information (e.g. ELIXIR). These efforts position ESGI as a key infrastructure taking care of incoming biological samples and generating genome-related data.

Furthermore, ethical, legal and societal issues were addressed to provide guidelines to adequately handle sequence data of patients in general and to specifically manage ethical issues related to transnational access projects.

Research and development (R&D) activities of ESGI focussed on improving ancillary wet-lab tools to make sample preparation for (specific) applications more efficient and less costly. Further R&D activities concerned the implementation of newest sequencing technologies to keep pace with international standards and the development of applications required from the user community such as single-cell sequencing. Moreover, software tools were developed to cater specific applications of the user community. The developments of our R&D activities were disseminated within the infrastructure through networking activities (see above), and were made available to the external users by early application in transnational access projects.

Provision of transnational access was the key activity of ESGI. Incoming projects from external users reflected the diversity of current biological research involving large-scale production of sequence data and subsequent multi-layered data analysis. The institutes of ESGI offered multiple specialities, both technologically and scientifically, to respond to the needs of the users. Support provided by ESGI facilities included sample preparation for sequencing and/or genotyping at different scales, the sequencing and genotyping processes themselves, raw data analyses and whenever needed support in in-depth biological analyses.

Tight interaction with users and ESGI facilities ensured that sample qualities matched the requirements for performing access studies efficiently. By offering "hands-on" access, the external users gained the possibility to stay for a period of up to three months at a given facility, to get familiar with the methodologies and data analysis methods required for access projects. Experienced staff at the ESGI facilities performed most of the wet-lab experiments whereas data analyses were done either by users themselves or very often in close collaboration with ESGI partners.

Altogether, ESGI announced three calls for proposals of transnational access projects. Based on an agreed procedure for selecting proposals for transnational access according to the EU guidelines of the capacities program, projects were selected based on the recommendation of in general two external and one internal scientific referee. Selection criteria for a study included scientific excellence, the significance of public health relevance, the potential to generate the number of cases required for genetic analyses, and the maturity in the field of research.

As the budget for access covered all costs required for a sequencing or genotyping study, the users did not have to bring in any additional means for the specific experiments planned with ESGI. In addition to providing access to sequencing and genotyping facilities, whenever needed, the partner sites of ESGI contributed specific biological expertise to enable successful completion of the multi-layered analyses to make the best out of complex sequence data.

Management structure and procedures

The Project Coordinator at MPG ensured the smooth operation of the project and guaranteed that all efforts were focused towards the objectives of the project. With the help of the assisting Project Manager (at GABO:mi), the coordinator submitted all required progress reports, deliverables, financial statements to the European Commission, and was responsible for the proper use of funds and their transfers to participants.

The ESGI project office was based at the coordinator in Berlin and at GABO:mi in Munich. The Project Office at the Coordinator was concerned with the scientific management and the co-ordination of all research activities, whereas the Project Office at GABO:mi was responsible for administrative, financial and contractual management and the organisational co-ordination of the project activities.

The Project Governing Board (PGB) was in charge of the political and strategic orientation of the project and acted as the arbitration body. It met roughly once a year unless the interest of the project required intermediate meetings. The Steering Committee consisted of all work package leaders and the Coordinator and was in charge of monitoring all activities towards the objective of the project in order to deliver in due time and in frame of the budget. The Steering Committee met in conjunction with the PGB meetings in person and via telephone conferencing in between those meetings. Furthermore, ESGI was backed up by a Scientific Advisory Boards, whose members were consulted whenever needed for solving specific problems, for example in the context of streamlining transnational access projects.

Objectives of ESGI:

The European Sequencing and Genotyping Infrastructure (ESGI) project was a successful joint effort of leading European sequencing, genotyping and bioinformatics centres to defragment the European research capability in a key area of the life sciences. ESGI provided external users with top-level access facilities to pursue demanding large-scale DNA analyses using first-class technologies for a broad range of genetics and systems biology applications.

In particular, massively-parallel sequencing technologies were essential components to unravel basic gene-regulating molecular pathways and disease processes, as well as to gain a better understanding of the evolution of European populations and disease susceptibility by deciphering the impact of variation of our genomes on clinical phenotypes.

The summarised accomplished goals of ESGI comprised development of an integrated distributed infrastructure of 7 major genome research sites and 4 complementing partner sites that:

- Provided state-of-the-art sequencing and genotyping systems including bioinformatics support for excellent genetics and systems biology research in Europe, which resulted in 29 transnational access projects of external users leading to a number of high-impact studies.
- Provided an infrastructure to routinely sequence the complete human genome in a few hours for less than 1000 Euro.
- Provided a flexible sustainable entity with a diverse application range in genetics and systems biology.
- Provided support and training in genome research for scientists in the 27 EU member states to contribute to a modern, science- and technology-based economy and society.

1.3 Description of the main S&T results/foregrounds of ESGI

ESGI facilities successfully established various networking and research and development (R&D) activities to efficiently test, establish and harmonise newest methods for nucleic acids analysis and sample preparation. These activities optimised and defragmented research capacities at various European genetics and genomics research facilities, by avoiding inefficient parallel efforts. Based on the establishment of an integrated technological infrastructure, provision of transnational access to highly sophisticated sequence analysis technologies to external users was the major activity of ESGI. After 54 months a significant improvement in knowledge transfer from the 7 involved large genomics centres and 4 complementing institutions of ESGI was achieved including additionally up to 29 external users or scientific groups.

We here describe along the lines of the 11 work packages (WP) of ESGI the main results that were achieved by this infrastructure pilot project.

WP1 (Management)

The management WP was led by GABO:mi, a German SME with longstanding, hands-on expertise in the management of EU Framework projects. Acting as the right hand of the Coordinator (MPG) and as a permanent help-desk for all project participants, GABO:mi took on full responsibility for project management and controlling, contractual management, communication management and resources management, thus ensuring professional project administration. Scientific coordination and management of the scientific activities and their progress was in the experienced hands of the Coordinator and the team of MPG. They had been building up large, internationally widely used infrastructure facilities such as the RZPD and have been coordinating large-scale national and EU integrated projects for years and were thus well familiar with the requirements for the scientific management of ESGI.

WP1 worked closely with WP8 (Transnational Access) in the management of the transnational access activities. GABO:mi helped in setting up the three calls for proposals via the ESGI website and the dissemination of information on these calls through email lists and announcements at meetings. Together, WP1 and WP8 developed a Memorandum of Understanding (MoU), governing the main operating principles between the transnational access users and the ESGI infrastructures. WP1 also assisted the leaders of the seven infrastructures with their local administration of the accepted transnational access projects.

In addition, WP1 supported WP6 in its dissemination and training activities through the ESGI public website, the project flyers as well as through the ESGI Symposium on Functional Genomics and Metabolism Research.

Networking activities

ESGI made additional substantial efforts to join forces between research facilities with various backgrounds and working culture. Therefore, a number of interconnected networking activities were intensified in 6 different work packages (namely WP2-7).

WP2 (Harmonisation)

The goal of WP 2 was to harmonize the sequencing and SNP genotyping procedures between the laboratories that participate as partners in ESGI.

A highlight of this WP2 was the contribution to a systematic RNA-sequencing quality data analysis and resource as published in: 't Hoen et al., Nat Biotechnol. 2013 and Lappalainen et al, Nature 2013. Further standard procedures for sequencing and genotyping were already harmonized to a large extent, because the sequencing/genotyping laboratories followed in general the most current protocols from the manufacturers of the respective sequencing/genotyping instruments. However, a number of specific applications were not covered by these procedures. The consortium thus collected various protocols, in particular for nucleic acids samples preparation from varying biomaterials, which were in use among the ESGI partners and made them available to each ESGI partner to adapt voluntarily. Furthermore, in collaboration with WP5, a set of guidelines for sample collection and management, compliant with BBMRI-ERIC, has been produced.

The transfer of "second generation sequencing" to the third generation (defined as a method to sequence the human genome for less than 1000 Euro in a few hours) was achieved by implementation of Illumina HiSeq X Five and Ten instruments in combination with new versions of sequencing chemistry. These new instruments allowed ESGI partners to sequence several human genomes per day for less than 1.000 EUR. P3-WTSI and P9-UU are currently the two ESGI members that have so far installed and implemented HiSeqX in their sequencing pipe-lines. Partners P5-CNG and P7-PCB are striving to acquire HiSeqX systems as well. The data from the HiSeqX instruments has a similar format as that from preceding HiSeq sequencing instruments, which has resulted in a rather smooth transfer to the new platform as the data can be stored, processed and shared using existing tools.

However the rate of data output from these new instruments has increased roughly four-fold compared to previous HiSeq instrument versions. The increased data amounts places considerable stress on available data storage and file transfer tools. To alleviate this P3-WTSI and P9-UU are currently investigating further how to utilize other, more compressed, data formats (such as compressed reference alignment format, CRAM).

The benchmark study, where several labs were provided exactly the same sequence data and were using similar variant calling pipe-lines showed how important it is not only to use the same software, but also that harmonized software settings are used (Alioto et al. BioRxiv 2014).

The additional progress of “third generation sequencing”, meaning true single molecule sequencing, has been slower than expected, although some partners within the ESGI consortium have invested in instruments from Pacific Biosciences. Standard operating procedures for PacBio instruments, laboratory information management system software and training of personnel have been accomplished. ESGI partner P9-UU has transferred its HPV-diagnostics from traditional Sanger sequencing to the PacBio platform, and P3-WTSI has used the PacBio instrument to sequence 3000 bacterial genomes for research purposes.

WP3 (Data storage and access)

This work package provided data management functions to the ESGI wet-lab facilities through a series of deliverables focused on data format standards, interfaces, submission and distribution pipelines, integration, security and harmonization. The data management domains and functions undertaken in this work package are described below. Owing to the rapid evolution of sequence analysis methods, solutions of data storage have to keep pace with these developments (Figure 1).

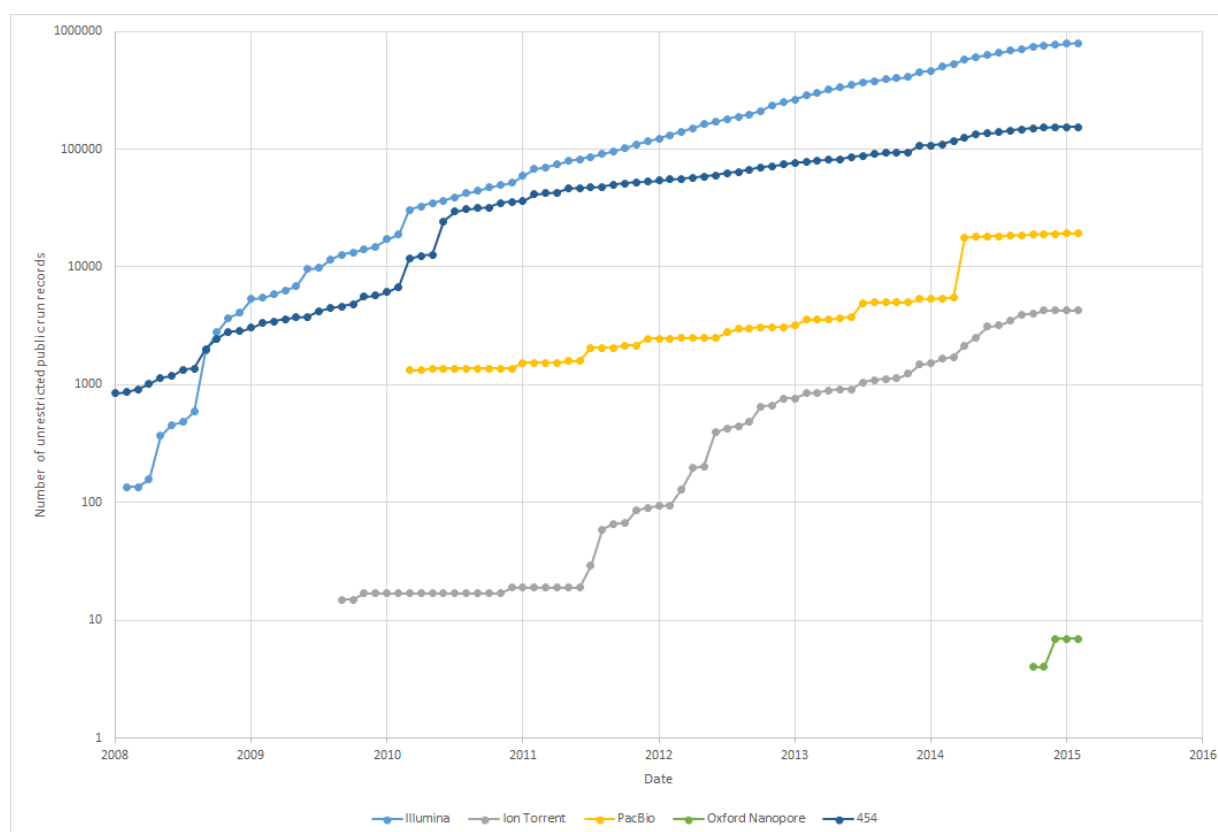


Figure 1: Growth in new genome sequencing technologies (based on Illumina, Ion Torrent, Pacific Biosciences, Oxford Nanopore and 454 (Roche)) in the EMBL-EBI archives.

Primary data and meta-data standards

Primary data formats define the file formats that report the outputs of sequencing and analysis pipelines, and can occupy hundreds of megabytes per file and terabytes of data for large studies. These formats such as BAM, CRAM, and VCF were first established as part of efforts such as the 1000 Genomes project, and their use was further refined for production archiving purposes as part of the ESGI project.

The standardized collection of meta-data, identifying the experimental and analytical methods, as well as study design, morphological tissue and disease phenotypes is an often overlooked but critical to the long term utility of

archived data and thus to ensuring the impact of investment in data generation. Meta-data are the source of significant practical challenges in the collection of rich data from submitters, without complicating and slowing the submission process. Much of the effort in streamlining the submission process such as the WebIn checklist module is aimed at achieving the balance of ease of submission and richness of meta-data.

Sample meta-data integration

Description of sample characteristics, such as sex, disease phenotype, and tissue origin often spans multiple physical experimental assays and can be distributed over time given the common use of biobank and cell-line derived samples. A single sample may be assayed at the DNA, RNA, epigenetic, proteomic, and metabolomics levels using different methods and under different conditions. This scenario places the tracking of sample data as a key data integration point. The EMBL-EBI BioSamples database fills this integration role across the axis of samples, as described below.

Data Submission

The requirements of data submission of both primary and meta-data were key to practical implementation of the data management and archiving system in support of ESGI projects, with a focus on enhancing usability and automation.

Data Distribution

Data distribution can be considered both in terms of data discovery, identifying datasets of interested, and the operational requirements for data request and data transfer.

Security

Human data presents additional data security requirements implemented in the EGA project and reported on in deliverable 3.3 and elsewhere.

Harmonization

Data formats and processes for ESGI were designed with interoperability as a primary requirement, with preferences to the examination and refinement of existing standards of practice. Key interactions with data submitter and international partners resulted in a system that is now a foundation for continued global efforts at harmonization of data and meta-data standards.

Operationally, tasks and deliverables were arranged to build on one another, while also delivering early critical functionality to ESGI data generating partners followed by iterative enhancements throughout the project. Feedback based of ESGI workshops, meetings, and interactions through the ENA and EGA help desks iteratively refined the usability requirements between the data archives and ESGI data submitters.

Primary data formats

Work package 3 sought to build upon existing primary file formats, defining recommendation and in some cases requirements which allow these formats to serve as a basis for a long term scientific data archive. For example, the explicit collection of a globally unique INSDC accession for a reference sequence, a submission requirement for BAM and VCF to the ENA and EGA and an inherent part of the reference based CRAM compression, ensures the long term utility and cross study comparability of archived data, beyond the file format definition itself.

The terabyte to petabyte scale of high throughput sequence data require efficient data management strategies during submission and delivery of data to users, and benefit from data compression strategies such as those implemented by CRAM.

A summary of primary data formats for ENA and EGA can be found in the ESGI D3.2 report and at <http://www.ebi.ac.uk/ena/submit/read-file-formats>.

Simplifying Meta-data submission with checklists

A key innovation developed in part during the ESGI project is the use of domain area specific checklists to simplify submission in way that captures the richness of meta-data specific to a given domain area, but without complicating submission with irrelevant options for all submitters. Further information about checklists can be found: <http://www.ebi.ac.uk/ena/submit/checklists>.

BioSample database

As described previously, sample information is a critical point of integration across multiple data collection experiments over the same sample or cell line. The BioSample database described here: <https://www.ebi.ac.uk/biosamples/> is now a critical part of the archiving system of both the ENA and EGA and serves this sample based cross study integration role. Further details were provided in report D3.5.

Data Submission

Streamlining the submission process using both interactive GUI and programmatic REST APIs to serve the ESGI data generation facilities was a key goal of this work package. Constructive feedback over the course of the project led to fully automated pipelines for data submission and error reporting. Further, facilities were put in place to better manage the multiple sources of information, for example sequence data from a sequencing core facility and meta-data from a scientific group, which may come from different points in within a submitting institution. The data submission process at the ENA and EGA is described at: <http://www.ebi.ac.uk/ena/submit> and <https://www.ebi.ac.uk/ega/submission>.

Data Discovery and Distribution

Data discovery has been enhanced through improved programmatic and interactive interfaces to the ENA described at: <http://www.ebi.ac.uk/ena/browse/programmatic-access> as well as through the collaboration in the EGA project between EMBL-EBI and the CRG Spain to provide a REST API based EGA interface for meta-data discovery. The EGA has moved to an improved internal modular architecture for delivery of secure download services described here: https://www.ebi.ac.uk/ega/about/your_EGA_account/download_streaming_client. Specific efforts to present control dataset were described in report D3.4.

The ESGI project has spanned a critical time period during the maturation of next-generation sequencing technologies and the associated data management infrastructure, and has helped to establish a set of data formats, interfaces, and data archiving and distribution practices which have led to the formation of global standards. Critical to the continued efforts at harmonization is interaction of EMBL-EBI with the NCBI in the United States and DDBJ in Japan in the INSDC consortium, which now includes the common next-generation meta-data model refined and applied in the ESGI project.

The impact of consistently archived reference datasets is particularly acute for the study of rare inherited diseases or cancers where experiments cannot readily be repeated. Publically available datasets are also key for the general goals of reproducible research and methods development.

High-throughput sequencing and the associated data generation and management challenges now touch all areas of the life sciences and impact major areas of social, economic and political concern such as agriculture and food security, biodiversity and the management of environmental and ecological impacts of human activity, assessing and adapting to climate change, biomedicine and the cost of delivery of clinical medicine and the management of public health risks such as the spread of anti-biotic resistant bacterial strains within clinical settings. All these evidence based application areas require reproducible and standard means to generate, store and reference results based on sequence data and serve as paths for impact from the work of the ESGI project. In the biomedical domain, the harmonization work undertaken in the ESGI project is continued through the participation of ESGI partners in the Global Alliance for Genomics and Health (GA4GH) (<http://genomicsandhealth.org/>).

WP4 (Analysis tools and computational support)

The main goals of this work package 4 were to exchange, update and harmonise computational tools that can be readily applied to perform the analyses of the wet-lab procedures required for transnational access projects.

To allow easy access to this resource a website has been set up, where transnational access users as well as the wider research community can find open access analysis tools for a variety of applications.

A further goal of WP4 consisted in harmonising the handling of biological information (to be generated by ESGI in addition to external users) with the European ELIXIR infrastructure.

The WP4 activity was focused on providing our integrated infrastructure with computational and bioinformatics support for the methods that are being harmonised through the activities of WP2 and which will be extensively applied in WP 8 (transnational access for external users). Sequencing technology is a fast moving field and there are ongoing developments to improve the analysis of the data. This is both done in large research centres which have established computational, bioinformatics and statistics support, and in smaller research groups.

We collated various bioinformatics tools that can be used for High Throughput Sequencing Data Analysis in order to allow the Transnational access users as well as other smaller research groups to use them.

Information and links to all the tools can be found under: <http://www.sanger.ac.uk/resources/software/esgi/>

The screenshot shows the website for the European Sequencing and Genotyping Infrastructure (ESGI) project. The page is titled "ESGI High throughput sequencing analysis tools". It describes the project as a collaborative effort funded by the European Union FP7 programme, involving a network of leading European genome research centres. The site provides information about bioinformatics and computational tools available for the analysis of high-throughput genomic data. A navigation menu includes tabs for Home, Research, Scientific resources, Work & study, and About us. Below the main content, there are tabs for various data analysis tools: Formats, Alignment, Assembly, Variant calling, Imputation, Structural variation, and RNAseq/ChIPseq. A "Related links" section points to the ESGI project site.

Figure 2: Screenshot - Sequencing tools

These computational tools have been used by the Transnational Access users, either by downloading the tools and running the analysis themselves or by the bioinformatics and statistical genetics experts involved in providing the infrastructure.

The public area of the website of the infrastructure hosts a download section, documentation, and a series of tabbed pages organising information about the tools and services. In this way it is anticipated that the site will be a useful resource to the wider research community.

The use of these High Throughput Sequencing Data Analysis tools can be automated by using pipeline management software, which allows the user to define a workflow and then push data through the resulting pipeline. A pipeline manager ensures that all the tools in the pipeline get run successfully, spreading the workload over a computer cluster. VRPipe (<https://github.com/VertebrateResequencing/vr-pipe/wiki>) was used at the Sanger Institute to do the bulk of the data processing for e.g. the 1000 Genomes Project, and can be used to run any of the software packages listed on the ESGI website (<http://www.sanger.ac.uk/resources/software/esgi/>).

To enable those centres with limited local computing infrastructure to carry out large-scale analysis, we adapted VRPipe so that it can now also be used in Amazon's cloud. A detailed guide for using VRPipe in Amazon's cloud is available under <https://github.com/VertebrateResequencing/vr-pipe/wiki/AWS> (D4.2)

The new sequencing technologies allow large quantities of sequencing data to be generated and bioinformatics is increasingly important. Also larger samples sets are required to even identify rare genomic events which have an effect on health or disease. To achieve suitable data set sizes it may be required to merge data from various sample collections. This will also help to make the most use of the data and the money invested. Standardising formats and data access is a valuable tool to achieve this. The sequencing centres involved in the ESGI project have submitted their sequencing data generated for the transnational access users to the European Nucleotide Archive (ENA) and European Genome-phenome Archive (EGA) housed at EMBL-EBI (partner 4). This will help to fulfil several requirements. The data generated within the ESGI project can be used by the wider scientific community and is of value in various projects and publishers require data to be made available. As part of the ESGI project the ELIXIR initiative has ensured development of new standards and formats have been developed for archiving new types of sequence data, such as epigenomic data and primary data from third generation sequencing platforms such as Pacific Biosciences. The data access needs of the ESGI project have supported the creation of a robust and extensible system for data submission, archiving, and data access.

Amongst others, we further provided improvements to already existing tools as well as newly developed tools for the analysis of data from various sequencing applications. The new or improved tools are PiBase (variants), B-SOLANA (methylation analysis), SeqBuster (miRNA), PeSVFisher (variant calling) and GEM Mapper (RNAseq). The

structural variant calling tool, PeSVFiser has been improved, by adding functionality (e.g. option to analyse a continuous targeted region of the genome, or modifying the tool to improve the user's experience).

Moreover, owing to installation of VRPipe on Amazon Elastic Compute Cloud (Amazon EC2) and allowing running the next generation sequencing (NGS) data analysis pipeline on cloud resources. The analysis of sequencing data requires a lot of compute power and complex pipelines using multiple software steps. By providing a version of the VRPipe pipeline from the Wellcome Trust Sanger Institute on the cloud computing service, ESGI has enabled users from institutions across Europe to access state of the art software infrastructure for NGS data analysis.

Finally, the data access needs of the ESGI project have supported the creation of a robust and extensible system for data submission, archiving, and data access as described in this and previous deliverable reports. New genome technologies are supported through a combination of specific extensions and meta-data checklists as well as built-in flexibility which allows the submission of "native" file formats and optional user-defined attribute fields. Downstream data types continue to converge in the standard CRAM and VCF representation of alignment and variant calls where applicable, and these approaches are being extended to support additional information such as epigenomic data where it is produced in tandem with sequence data by new technologies.

WP5 (Networking to biobanking)

WP05 provided the interface between ESGI and leading international and European biobanking initiatives, such as BBMRI-ERIC, an ESFRI BMS Research Infrastructure.

In this workpackage, harmonized standards were developed relating to handling of biological samples, pre-analytical processing, and isolation, storage and quality control specifically of DNA and RNA. These standards took into account the quality requirements of the latest sequencing and genotyping technologies. Furthermore, appropriate procedures for the exchange of (human) biological samples and data, specifically sensitive medical data, were defined, that account for sensitive issues of privacy, ethical and legal issues and also for intellectual property rights in the extremely heterogeneous landscape of European legislation and ethical regulations. The criteria defined in this WP serve as guidelines for the access to samples and data from biobank networks, such as BBMRI. The work in WP05 was structured into four tasks that contribute to the objectives detailed above.

The first task addressed the quality issues of biological samples with specific focus on DNA and RNA quality. Interfacing ESGI and BBMRI-ERIC requires compatibility regarding scientific, ethical and technical requirements. Key principles for access are the primacy of European legal and ethical regulations, a clearly defined access policy, and a data management and protection policy.

In a report that also provided a set of guidelines for access projects (deliverable D5.01), the key quality criteria applying to biobanks containing human biological material and associated data were laid down. For all human biological material and associated data to be analysed by ESGI user projects their source and quality must be defined. On the one hand, this requires that this biological material and the associated data can be accessed in a transparent way. Specific quality management must be available, and the processes must comply with ethical and legal prerequisites; conditions that are generally fulfilled by biobanks. In particular, the OECD best practice guidelines for Global Biological Resource Centre Networks are to be adopted as the common standard regarding infrastructure, management, traceability, biosecurity and biosafety, data protection, provision of data (minimal/recommended datasets) and quality management.

In cooperation with the FP7 project SPIDIA, a feasibility study was performed which allowed to define further the requirements on sample quality with particular emphasis on sequencing. In this project we defined also templates for sample and data transfer and specified appropriate pre-analytical treatment of samples using advanced methods of sample stabilization or cryopreservation.

In a second major task (documented by deliverables D5.02, D5.04 – D5.06, D5.08), we elaborated the rules for mutual access to biological samples and phenotype data, as well as to the sequencing/genotyping infrastructure. Several deliverables provided the various items required in this process, such as templates for Confidential Disclosure Agreements (CDA) and Material Transfer Agreements (MTA) (D5.02), best-practice guidelines for sample management (D5.05, D5.08), and for data management, pooling and exchange (D5.06).

The consensus MTA and CDA templates were compiled after revising MTAs/CDAs from several ESGI partners. The consensus MTA was tested in a pilot test run that involved CRG and MUG for their practical applicability. This form was subsequently used for the access projects.

In three calls, 29 projects were selected for transnational access to the ESGI infrastructure. These projects were evaluated after the last call with respect to their current status of sample quality assurance (D5.08), using detailed questionnaires.

A dedicated workshop (D5.04) brought together ESGI participants and representatives from BBMRI and SPIDIA. In this workshop, major achievements of WP5, Tasks 1 and 2 were presented and discussed. The results contributed substantially to the 'ESGI best practice guidelines/recommendations for sample management' (D5.05).

Our third task (D5.03) dealt with the harmonization of phenotype data and their integration. The heterogeneity in structure and quality of medical data necessitates solutions for the extraction of medical data from records, data clean-up and their integration and harmonization. Moreover, solutions for access to data must be provided that are sufficiently detailed to allow their use in research while still guaranteeing privacy, e.g. avoiding re-identification of the patients described in anonymized records by combination. A recent solution to this problem is provided through specific algorithms for data aggregation (k-anonymity). In a report the current status of harmonization of terminology and data, of specific strategies to deal with complex data (visual analytics) and of maintaining privacy while allowing maximal use of data for research was summarized.

In particular, (1) through cooperation with BBMRI and BioMedBridges, harmonization of phenotype data was advanced, using common identifiers, meta-standards for samples, a registration and annotation service, and semantic standardization. (2) – Visual Analytics was applied to gene expression analysis, pathway analysis and clinical data, all of which are examples for data sets of extreme complexity. This approach was integrated into open source biobank software ('BIBBOX'). (3) – The advantage of the k-anonymity approach is to maintain privacy while allowing maximal usability of the data. This is achieved by allowing the user to define the level of aggregation of specific data fields that are required for the research purpose, trading in a higher degree of aggregation in other data fields of lesser importance. For practical applications, K-anonymity was incorporated into an open access tool ('OpenAnonymiser').

A fourth task (D5.07) was dedicated to the appropriate consideration of ethical and legal requirements that inevitably accrue in transnational exchange of biological samples and medical data. The 29 projects that were granted access by ESGI originated from 15 European countries with widely different ethical and legislative requirements. The projects were subjected to evaluation of their ethical impact and also of the overall ethical approach of ESGI (in WP7). This was also achieved through a panel of external reviewers. As a result of this evaluation it was possible to propose measures to improve the ethical assessment of innovative biomedical projects.

Overall, WP5 aimed at integrating biobanking with the analysis of samples and data, specifically BBMRI-ERIC, and ESGI. The major impact of this integration process is that it provides the basis for efficient and reliable sample and data sharing/exchange for the genetic investigation of new clinical and epidemiologic sample collections.

Through the access of 29 projects to the ESGI infrastructure the guidelines and tools that were established in this work package were thoroughly tested for their reliability, practicability and usefulness regarding sharing of samples and data. Several tools or templates were elaborated that facilitate this complex process, and that help maximize the use of the valuable resources provided by biobanks and the analysis of samples and data. The ethical evaluation of these projects also provided recommendations for the improvement of ethical assessment of EC funded innovative biomedical research.

WP6 (Dissemination and training)

ESGI aimed to create additional awareness amongst a broad range of scientists about the potential of genetics and genomics to elucidate complex molecular pathways and disease processes. Dissemination of scientific know-how was essential for progress in the field and to organise a competition of the best projects of external users to be selected for analyses at the infrastructure. Training workshops, online tools and other outreach activities formed the bulk of this work package. Aside from this, ESGI committed to keeping funding agencies and scientists within the field informed by its various activities.

Therefore, WP6 successfully completed its specific tasks and deliverables. Those included a public website established for information about the infrastructure consortium and launch of calls for proposals of transnational access projects. The ESGI web site (<http://www.esgi-infrastructure.eu/>) had been launched in March 2011. The site provides information about the aims of the project, the structure and the participating institutions. It also provided information for the three transnational calls for proposals (<http://www.esgi-infrastructure.eu/transnational-access/call/>) and the press release describing those calls (http://www.esgi-infrastructure.eu/fileadmin/websites/esgi/media/ESGI_PressRelease_20110314.pdf).

WP6 also delivered an ESGI brochure that had been distributed and was available via the consortium web site (http://www.esgi-infrastructure.eu/fileadmin/websites/esgi/media/ESGI_Flyer_complete.pdf). The brochure describes the participants, background and purpose of ESGI, and alerts scientists to the possibility of working with ESGI to obtain genotyping and sequencing services. Perhaps most importantly, it provides a link to the ESGI website for interested readers. In combination with the flyer, there was also a poster produced: <http://www.esgi->

infrastructure.eu/fileadmin/websites/esgi/media/Plakat_ESGI_A1_final.pdf. The poster and brochure were distributed by the participants of ESGI at various international meetings. The brochure was updated again in 2014.



Figure 3: Front page of the Brochure of ESGI.

As one of its core tasks, WP6 offered several workshops including

- “Large workshop 1”: the 4th Paris Workshop on Genomic Epidemiology that was held in Paris from 30/05/2011 – 01/06/2011. The web site, including the program can be found at <http://www.cng.fr/workshop2011/overview.html>. ESGI was a key sponsor of the meeting and the more than 200 delegates listened to a set of excellent talks including papers presented by ESGI participants (Ivo Gut, Harold Swerdlow, Sascha Sauer, Kurt Zatloukal, Ann-Christine Syvaenen, Anne Cambon-Thomsen).
- “Large workshop 2”: the ESGI-sponsored workshop on Genomic Epidemiology was held in Paris May 2-4 2013. The workshop info can be accessed on: http://innovationcenter.netne.net/paris_workshop/sponsors.html. Five ESGI speakers from member institutes presented: Roderic Guigó,, Ivo Gut , Sascha Sauer , Hans Lehrach , Stefan Schreiber.
- “Workshop 1”: a training workshop dedicated to “next-generation sequencing” (NGS): The training course consisted of a mix of theoretical and practical elements related to next-generation sequencing, especially Solexa sequencing as performed on the Illumina platform. The course was held jointly by the Wellcome Trust Advanced Courses and ESGI. The site was the Wellcome Trust Sanger Institute in Hinxton, UK from the 2nd - 10th October 2011. See: <http://www.wellcome.ac.uk/Education-resources/Courses-and-conferences/Advanced-Courses-and-Scientific-Conferences/Past-events/Advanced-courses/index.htm>.

The course covered many aspects both practical and theoretical of:

- Next-generation sequencing library preparation
- Sequencing on the Illumina HiSeq platform
- 3rd-generation and benchtop sequencing technologies
- Bioinformatics
- Many diverse applications of DNA sequencing from Archaeology to Modern Genomic Epidemiology.

- “Workshop 2”: a second training workshop titled “ESGI Data Flow Workshop 2012” (February 8th and 9th 2012, EMBL-EBI, Hinxton, Cambridge) that provided training covering submission and retrieval services provided by the EBI’s European Nucleotide Archive (ENA; <http://www.ebi.ac.uk/ena/>) for raw and early analysis data from next generation nucleotide sequencing platforms. In addition, it covered emerging technology for the efficient compression of these data. Students were expected to become familiar with submission and retrieval formats for both metadata and data, submission pipelines and data retrieval using browser and webservice methods, and the use of compressed formats in data processing and analysis. More details are available here: <http://www.ebi.ac.uk/training/course/ebi-affiliated-course-esgi-data-flow-workshop>
- “Workshop 3”: the ESGI Symposium on Functional Genomics and Metabolism Research (third ESGI workshop), held on March 21st and 22nd in Berlin, Germany (<http://www.esgi-infrastructure.eu/meetings/symposium2013/>). Numerous ESGI members participated including Sascha Sauer, Richard Durbin (WTSI), Ivo Gut (CNAG), Ann-Christine Syvaenen (UU), Andre Franke (CAU).
- “Workshop 4”: a fourth training workshop titled “Biological Interpretation of Next-Generation Sequencing Data”, held at EBI in Hinxton from December 2nd to 6th, 2013. Topics included:
 - Introduction to NGS analysis and ChIP-Seq
 - RNA-seq analysis
 - Epigenetic data integration
 - Gene expression data integration
 - Integration across genomes

In summary this has been a very successful work package enabling the training and education of several hundred scientists from a wide variety of institutions across the European Union. Knowledge of NGS (next generation sequencing) approaches and applications has been disseminated to eminent scientists as well as more junior scientists that will form the next generation of researchers. The training activities of ESGI have been timed perfectly to coincide with the expansion of the use of NGS in the clinical medicine. This in particular will have particular benefit to public health within the EU.

Furthermore since the market for NGS goods and services is totally tens of billions of Euros per annum, this is bound to have profound economic benefits right across the region.

WP7 (Ethical, legal and societal issues)

The scientific activities of the ESGI project imply a number of specific ethical and legal issues ranging from the protection of confidentiality and privacy, consent to the use of human samples, feedback of genetic and medical information, ethical sourcing of the samples, consideration for the welfare of non-human animals and legal compliance with national and international standards for human, animal and plant research. Work Package 7 has therefore been established with the aim of providing ethical guidance and guaranteeing ethically defensible conduct of research and respect for relevant legal regulations. Moreover, WP7 has had the role of fostering reflection and providing ethical and legal training within the project, as well as a follow up of ethical and legal regulation and public debates on topics relevant for ESGI.

Given the varied nature of the projects that gained access to the ESGI platform for sequencing and analysis, spanning from human, animal, micro-organism and plant genetics, WP7 set up specific policies for each of these domains in order to fulfil its main objective: the development and implementation of a comprehensive framework for the ethical governance of the ESGI infrastructure.

WP7 has taken care of the following tasks:

1. Ever since the beginning of the project, WP7 has been surveying the ethical, legal and societal issues that directly bear on the activities of the infrastructure, as well as those that apply more generally to the field of genotyping and next-generation sequencing (D7.2, D7.4, D7.6). This activity included, closely following the revision of the European Data Protection Directive, while keeping the partners informed of the changes that may have been relevant to the consortium. Also the revision of the Council of Europe recommendation for the research use of biological elements of human origin was followed and a contribution forwarded.
2. We conceived of a template for the preliminary assessment of access projects as regards to their ethical, legal and social components (D7.1).
3. We ensured that all access projects complied with the established ethical and legal requirements (D7.1, D7.5, D7.7). In particular we assessed the ethical templates from these projects and, based on a case-by-case analysis, we assigned them for external ethical review to competent experts in the field of research ethics. At

times this implied undertaking personal interaction with principal investigators in order to provide advice, ask for necessary clarification and to solve issues with missing documentation. All the projects were eventually approved from an ethical point of view.

4. The ethical framework for ESGI has been made fully available and explained to the rest of the consortium through deliverable D7.5 and D7.7.
5. We managed to coordinate our activities with those of relevantly similar projects so as to reduce duplication of work and to be sure to be following best practices in the field (D7.3).
6. The analyses produced for the project as to the ethical, legal and social issues in next-generation sequencing (especially D7. 2, D7.4, D7.5, D7.6, D7.7) were used as the basis for internal training activities as well as for education purposes outside of the project.
7. We analysed data sharing practices in next-generation sequencing genomics especially as to professional attitudes towards the handling of large scale genetic data produced in the course of NGS projects.
8. We have provided proposals for the sustainability of the ethical framework of ESGI (D7.8).
9. We have proposed lines of research for further collaboration among ESGI partners on ethical issues in NGS.

WP7 has successfully accomplished its tasks. First and foremost we ensured the development and implementation of an ethical framework that proved sufficiently comprehensive, workable and effective throughout the duration of the project. In this respect, the procedures established for ESGI can easily represent a standard for the future and therefore can be regarded as a reliable framework for collaborative projects in NGS featuring complex ethical and practical issues. The ESGI framework and templates were provided as basis or examples for others; it can especially be mentioned their use by BBMRI-ERIC for the establishment of the ethics check procedure of this infrastructure.

Moreover, WP7's work on the analysis of ethical, legal and social implications of next-generation sequencing occasioned participation in the international ethical debate on crucial issues such as the feedback of research results and incidental findings. Also, this enabled WP7 to participate in the development of the European Society for Human Genetics recommendations concerning whole genome and whole exome sequencing and the issue of incidental findings as well as to collaborate with further recommendations in collaboration with another FP7 project 3Gb-TEST.

WP7 managed to expand the scope of our analyses to newly emerging issues for genetics and genomics and we succeeded in raising the level of awareness – both within the consortium and through its broader network of collaborators – regarding pressing societal issues that, normally, tend to be ignored. In this respect, deliverable D7.6 (and related activities), represented an opportunity to expand the scope of science's engagement with potentially critical issues arising in the immediate proximity of research activities.

WP7 managed to coordinate efficiently with other European and International projects presenting similar ethical issues as ESGI. This ensured that ESGI could benefit from emerging best practice in its domain of application.

The impact of the work performed by an Ethics work-package can primarily be assessed by looking at its effects on the activity of the scientific research. Thanks to the carefully crafted procedures for ethical assessment set up under the coordination of WP7, 29 transnational access projects gained access to the sequencing and analysis services offered by ESGI partners. Assessment procedures were followed scrupulously by all access projects. The procedures and the availability of WP7 to directly interact with principal investigators offered the researchers the opportunity to focus on their scientific aim – knowing that ethical and legal compliance was being properly advised by a dedicated team in coordination with international experts. In this respect, from the point of view of ethical governance, ESGI certainly represents a model of good practice – as testified by the absence of any ethical “accident“ for the entire duration of the project.

On a more theoretical side, the activities of WP7 aimed at surveying, clarifying and analysing ethical and societal issues in next-generation sequencing, as well as legal issues in the domain of high-throughput science, will constitute the basis for further research. In particular, the activity of WP7 in this area highlighted that many emerging issues in next-generation sequencing are far from settled at the level of moral theory. The theoretical panorama in this domain appears to be still under construction. This may create sub-optimal levels of indeterminacy that can, eventually, result in overly burdensome governance framework or, at the other side of the normative spectrum, can lead to overlook important ethical trade-offs. For this reason, hands-on experience on a large scale project such as ESGI, coupled with the systematic exploration of the theoretical issues at stake, has represented an opportunity to gear ethical reflection towards more accurate and fine-grained understanding of the challenges that lay ahead the scientific, technological and medical development of genome sequencing.

Transnational access

WP8 (Transnational access)

Transnational access was the core activity of ESGI. ESGI provided direct access or alternatively remote services, including in particular application of high-throughput sequencing technologies of the second or third generation. The demands from users for medium- and high-multiplex genotyping methods were instead very low. In general, ESGI mainly offered support to external users for the following 3 main steps of high-throughput sequencing:

- i. Preparation of nucleic acids samples for sequencing (and genotyping) applications. This included amongst others library preparation and genomic enrichment/pull-down of genomic loci for re-sequencing.
- ii. Offering sequencing (and genotyping) from the high-throughput instruments at the facilities. During this reporting period most facilities applied the HiSeq2000/2500 sequencers from Illumina.
- iii. Bioinformatics analyses. This part included raw data and higher-level. According to the specific needs of the users, ESGI facilities offered further bioinformatics support that is tailored to the diverse projects. In particular, this type of support turned out to become crucial for many projects.

Description of the publicity concerning the new opportunities for access:

ESGI organised altogether 3 open calls for transnational access project proposals and provided information packages on the rules and objectives of ESGI via the project website (www.esgi-infrastructure.eu). For each call, ESGI defined scientific topics to respond best to the broad research community and expertise at ESGI facilities and to streamline the selection process.

Calls for proposal and further information was mainly distributed via the national contact agencies of the member and associated states and by using distribution channels established at the different ESGI facilities such as international and national conferences in genetics and genome research.

Interested potential users could submit project proposals directly via the ESGI website. Questions with regard to project proposals could be sent via a dedicated email address, and were in general answered within a day by the ESGI coordinator (transnational access work package leader) or the ESGI project manager.

After decision, successful applicants were informed by email. Non-successful applicants were also informed about the decision by email around the same time as the successful applicants. ESGI provided more detailed feedback to those users who explicitly asked for that, but this was rarely the case. Successful applicants had a telephone conference with the coordinator of ESGI, Dr. Sascha Sauer, to learn more about ESGI services, potential streamlining of studies according to the comments of referees, and further specific topics such as ethical issues related to the user projects. Afterwards the users contacted the ESGI facility leaders to discuss practical issues of the projects. The ESGI project coordinator signed with each of the external users and the responsible facility leader of the dedicated installation a Memorandum of Understanding to set rules for the scientific interaction in the frame of Transnational Access projects.

The long-term availability of ESGI partners for users was important to ensure publication of results (e.g. to address scientific criticism of referees of peer-reviewed journals to improve the quality of manuscripts of the users). In general the average access time at ESGI facilities did not exceed three months; but mid-to long term opportunity to access ESGI platforms was almost for every user required to ensure completion of experiments and success of user projects. This aspect may distinguish a genomics facility from other infrastructures, to ensure required support of biologists and medical researchers who generally apply iterative experimental approaches.

Description of the selection procedure

Each proposal was in general reviewed and scored by two external referees. ESGI gathered a panel of internationally well-known experts, whose diverse expertise – comprising genetic variation analyses, population genetics, functional genomics, disease-oriented genetics and genomics, etc. - corresponded with the different scientific topics of our calls. In addition, per proposal one internal ESGI referee contributed to the selection process.

Proposals that were selected based on the criteria defined in the grant agreement: 1. general scientific merit of the proposal and potential impact of the proposed project; 2. quality of clinical and other phenotypic data; 3. accessibility of biological samples that can be transformed into high quality nucleic acids; 4. potential of the collection to advance current knowledge of genetics of disease; 5. ethical approval and consent to exchange nucleic acids samples; and 6. agreement on data sharing policy. Scientific excellence and impact were the most important criteria for the selection.

ESGI steering and project selection committee members decided to allocate projects for about one third of the transnational access budget per call. According to this number, we defined an average score of undisputed projects, which were allocated to different ESGI sites according to the strengths of specific methodology and various expertise at ESGI partner sites (e.g. population genetics projects were allocated to the WTSI Sanger Institute).

Selected project proposals were further evaluated for compliance to ethical standards by one external and one ESGI-internal expert (Dr. Anne Cambon-Thomsen). Feedback was provided to make users and facility leaders aware of any potential problems. Many access projects dealt with basic functional genomics studies in cell culture or animal models or studies with patient DNA samples that have already been approved in other biological or genetics studies, which rendered the ethical check in many cases unproblematic.

Altogether ESGI could support 29 projects, selected from 103 eligible applications.

Significant results

ESGI has successfully finished the 29 selected projects with a broad range of applications and scientific focus. The geographic distribution of external users is shown in Figure 4.



Figure 4: Geographic distribution of 29 projects selected for transnational access

Initial list of studies of external users of ESGI that were already published by the end of the project:

[Homozygous loss-of-function variants in European cosmopolitan and isolate populations.](#)

Kaiser VB, Svinti V, Prendergast JG, Chau YY, Campbell A, Patarcic I, Barroso I, Joshi PK, Hastie ND, Miljkovic A, Taylor MS; Generation Scotland; UK10K, Enroth S, Memari Y, Kolb-Kokocinski A, Wright AF, Gyllensten U, Durbin R, Rudan I, Campbell H, Polašek O, Johansson Å, Sauer S, Porteous DJ, Fraser RM, Drake C, Vitart V, Hayward C, Semple CA, Wilson JF.

Hum Mol Genet. 2015 Oct 1;24(19):5464-74. doi: 10.1093/hmg/ddv272. Epub 2015 Jul 14.

PMID: 26173456

[Deficiency of ECHS1 causes mitochondrial encephalopathy with cardiac involvement.](#)

Haack TB, Jackson CB, Murayama K, Kremer LS, Schaller A, Kotzaeridou U, de Vries MC, Schottmann G, Santra S, Büchner B, Wieland T, Graf E, Freisinger P, Eggimann S, Ohtake A, Okazaki Y, Kohda M, Kishita Y, Tokuzawa Y, Sauer S, Memari Y, Kolb-Kokocinski A, Durbin R, Hasselmann O, Cremer K, Albrecht B, Wiczorek D, Engels H, Hahn D, Zink AM, Alston CL, Taylor RW, Rodenburg RJ, Trollmann R, Sperl W, Strom TM, Hoffmann GF, Mayr JA, Meitinger T, Bolognini R, Schuelke M, Nuoffer JM, Kölker S, Prokisch H, Klopstock T.

Ann Clin Transl Neurol. 2015 May;2(5):492-509. doi: 10.1002/acn3.189. Epub 2015 Mar 13.

PMID: 26000322

[Deep sequencing of the *Trypanosoma cruzi* GP63 surface proteases reveals diversity and diversifying selection among chronic and congenital Chagas disease patients.](#)

Llewellyn MS, Messenger LA, Luquetti AO, Garcia L, Torrico F, Tavares SB, Cheaib B, Derome N, Delepine M, Baulard C, Deleuze JF, Sauer S, Miles MA.

PLoS Negl Trop Dis. 2015 Apr 7;9(4):e0003458. doi: 10.1371/journal.pntd.0003458. eCollection 2015 Apr.

PMID: 25849488

[Immunofluorescence Analysis and Diagnosis of Primary Ciliary Dyskinesia with Radial Spoke Defects.](#)

Frommer A, Hjej R, Loges NT, Edelbusch C, Jahnke C, Raidt J, Werner C, Wallmeier J, Große-Onnebrink J, Olbrich H, Cindrić S, Jaspers M, Boon M, Memari Y, Durbin R, Kolb-Kokocinski A, Sauer S, Marthin JK, Nielsen KG, Amirav I, Elias N, Eitan K, Shoseyov D, Haeffner K, Omran H.

Am J Respir Cell Mol Biol. 2015 Mar 19. [Epub ahead of print]

PMID: 25789548

[Development of a high-resolution NGS-based HLA-typing and analysis pipeline.](#)

Wittig M, Anmarkrud JA, Kässens JC, Koch S, Forster M, Ellinghaus E, Hov JR, Sauer S, Schimpler M, Ziemann M, Görg S, Jacob F, Karlsen TH, Franke A.

Nucleic Acids Res. 2015 Jun 23;43(11):e70. doi: 10.1093/nar/gkv184. Epub 2015 Mar 9.

PMID: 25753671

[Transcription-dependent generation of a specialized chromatin structure at the TCR \$\beta\$ locus.](#)

Zacarias-Cabeza J, Belhocine M, Vanhille L, Cauchy P, Koch F, Pekowska A, Fenouil R, Bergon A, Gut M, Gut I, Eick D, Imbert J, Ferrier P, Andrau JC, Spicuglia S.

J Immunol. 2015 Apr 1;194(7):3432-43. doi: 10.4049/jimmunol.1400789. Epub 2015 Mar 2.

PMID: 25732733

[Site- and allele-specific polycomb dysregulation in T-cell leukaemia.](#)

Navarro JM, Touzart A, Pradel LC, Loosveld M, Koubi M, Fenouil R, Le Noir S, Maqbool MA, Morgado E, Gregoire C, Jaeger S, Mamessier E, Pignon C, Hacein-Bey-Abina S, Malissen B, Gut M, Gut IG, Dombret H, Macintyre EA, Howe SJ, Gaspar HB, Thrasher AJ, Ifrah N, Payet-Bornet D, Duprez E, Andrau JC, Asnafi V, Nadel B.

Nat Commun. 2015 Jan 23;6:6094. doi: 10.1038/ncomms7094.

PMID: 25615415

[High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis.](#)

Goyette P, Boucher G, Mallon D, Ellinghaus E, Jostins L, Huang H, Ripke S, Gusareva ES, Annese V, Hauser SL, Oksenberg JR, Thomsen I, Leslie S; International Inflammatory Bowel Disease Genetics Consortium; Australia and New Zealand IBDGC; Belgium IBD Genetics Consortium; Italian Group for IBD Genetic Consortium; NIDDK Inflammatory Bowel Disease Genetics Consortium; United Kingdom IBDGC; Wellcome Trust Case Control Consortium; Quebec IBD Genetics Consortium, Daly MJ, Van Steen K, Duerr RH, Barrett JC, McGovern DP, Schumm LP, Traherne JA, Carrington MN, Kosmoliaptsis V, Karlsen TH, Franke A, Rioux JD.

Nat Genet. 2015 Feb;47(2):172-9. doi: 10.1038/ng.3176. Epub 2015 Jan 5.

PMID: 25559196

[Refinement of the MHC risk map in a scandinavian primary sclerosing cholangitis population.](#)

Næss S, Lie BA, Melum E, Olsson M, Hov JR, Croucher PJ, Hampe J, Thorsby E, Bergquist A, Traherne JA, Schrupf E, Boberg KM, Schreiber S, Franke A, Karlsen TH.

PLoS One. 2014 Dec 18;9(12):e114486. doi: 10.1371/journal.pone.0114486. eCollection 2014.

PMID: 25521205

[PRMT1 and PRMT8 regulate retinoic acid-dependent neuronal differentiation with implications to neuropathology.](#)

Simandi Z, Czipa E, Horvath A, Koszeghy A, Bordas C, Póliska S, Juhász I, Imre L, Szabó G, Dezsó B, Barta E, Sauer S, Karolyi K, Kovacs I, Hutóczki G, Bognár L, Klekner Á, Szucs P, Bálint BL, Nagy L.

Stem Cells. 2015 Mar;33(3):726-41. doi: 10.1002/stem.1894.

PMID: 25388207

[Efficient application of next-generation sequencing for the diagnosis of rare genetic syndromes.](#)

Madrigal I, Alvarez-Mora MI, Karlberg O, Rodríguez-Revenga L, Elurbe DM, Rabionet R, Mur A, Pie J, Ballesta F, Sauer S, Syvänen AC, Milà M.

J Clin Pathol. 2014 Dec;67(12):1099-103. doi: 10.1136/jclinpath-2014-202537. Epub 2014 Sep 30.

PMID: 25271213

[Divergent transcription is associated with promoters of transcriptional regulators.](#)

Lepoivre C, Belhocine M, Bergon A, Griffon A, Yammine M, Vanhille L, Zacarias-Cabeza J, Garibal MA, Koch F, Maqbool MA, Fenouil R, Loriod B, Holota H, Gut M, Gut I, Imbert J, Andrau JC, Puthier D, Spicuglia S.

BMC Genomics. 2013 Dec 23;14:914. doi: 10.1186/1471-2164-14-914.

PMID: 24365181

[Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis.](#)

Liu JZ, Hov JR, Folseraas T, Ellinghaus E, Rushbrook SM, Doncheva NT, Andreassen OA, Weersma RK, Weismüller TJ, Eksteen B, Invernizzi P, Hirschfield GM, Gotthardt DN, Pares A, Ellinghaus D, Shah T, Juran BD, Milkiewicz P, Rust C, Schramm C, Müller T, Srivastava B, Dalekos G, Nöthen MM, Herms S, Winkelmann J, Mitrovic M, Braun F, Ponsoen CY, Croucher PJ, Sterneck M, Teufel A, Mason AL, Saarela J, Leppä V, Dorfman R, Alvaro D, Floreani A, Onengut-Gumuscu S, Rich SS, Thompson WK, Schork AJ, Næss S, Thomsen I, Mayr G, König IR, Hveem K, Cleynen I, Gutierrez-Achury J, Ricaño-Ponce I, van Heel D, Björnsson E, Sandford RN, Durie PR, Melum E, Vatn MH, Silverberg MS, Duerr RH, Padyukov L, Brand S, Sans M, Annese V, Achkar JP, Boberg KM, Marschall HU, Chazouillères O, Bowlus CL, Wijmenga C, Schrupf E, Vermeire S, Albrecht M; UK-PSCSC Consortium; International IBD Genetics Consortium, Rioux JD, Alexander G, Bergquist A, Cho J, Schreiber S, Manns MP, Färkkilä M, Dale AM, Chapman RW, Lazaridis KN; International PSC Study Group, Franke A, Anderson CA, Karlsen TH.

Nat Genet. 2013 Jun;45(6):670-5. doi: 10.1038/ng.2616. Epub 2013 Apr 21.

PMID: 23603763

[CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters.](#)

Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I, Andrau JC.

Genome Res. 2012 Dec;22(12):2399-408. doi: 10.1101/gr.138776.112. Epub 2012 Oct 25.

PMID: 23100115

[Threonine-4 of mammalian RNA polymerase II CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation.](#)

Hintermair C, Heidemann M, Koch F, Descostes N, Gut M, Gut I, Fenouil R, Ferrier P, Flatley A, Kremmer E, Chapman RD, Andrau JC, Eick D.

EMBO J. 2012 Jun 13;31(12):2784-97. doi: 10.1038/emboj.2012.123. Epub 2012 May 1.

PMID: 22549466

A number of studies are currently submitted for peer-reviewing to international scientific journals, are being prepared for publication, or still require (ongoing) experiments that go beyond the scope of support provided by ESGI.

During the course of the ESGI project it became apparent that many external users required not only access to wet-lab technologies but also needed access to substantial bioinformatics infrastructures and support in analyses of

large-scale sequence data. This issue will become even more important with the recent developments of sequencing technologies that produce even more impressive amounts of data in a short time.

Beyond the high scientific impact expected from the transnational projects, this core WP8 allowed the distributed centres of ESGI to set up a sustainable framework for future support of an expected increasing number of external users.

Research and development

WP9-11 included research and development activities to install the most powerful technologies in the field of genome research and to improve the technologies applied for Transnational Access. The results of these activities at the various sites of ESGI were distributed via the networking activities (mainly WP 2, 4, 6).

WP9 (Improvement of existing methods) Objectives

As the title of this work package says it focuses on taking technologies and methods in sequence analysis and improve them so that they perform reliably, are robust and can be taken into production. We had three main objectives and a series of ancillary objectives. This activity is important in the context that many procedures and methods that are published are developed with the objective to solve a problem once. As such they are not refined for robustness, performance and integration into a routine environment. A further point is that many methods have a solution, but upon application of the methods it becomes clear that quality is below what potentially could be achieved. The main targets for this work package were on methodology for expression quantitative trait loci (eQTL) analysis, targeted capture and copy number variant (CNV) analysis. In addition we wanted to improve sample preparation procedures (without PCR, for lower input amounts and for material extracted from FFPE samples), automate sample preparation procedures and improve on analysis procedures for sequencing data. The different objectives overlapped with competence held by different ones of the partners. This seeming redundancy allowed assessing and complementing approaches.

A number of improved methods for sequence analysis were achieved. Firstly, a suite of tools has been developed for eQTL analysis. The tools are tuned for array-based expression analysis and for RNA sequencing-based analysis (Almlöf et al. PLoS One 2012, Almlöf et al. PLoS One 2014, Adoue et al. Molecular Systems Biology 2014). This knowhow was used for the analysis in Lappalainen et al. Nature 2013.

The benchmarking and implementation of tools for genome enrichment were a further achievement. In general the interest in these technologies has been substantial, mainly because contrary to general expectation the cost of whole genome sequencing remained constant for several years and did not sink to the 1000 Euro genome as expected (this finally happened in the last year). Due to this in particular exome sequencing became very popular and several companies started offering kits. This required testing of different products. Also different enrichment technologies were tested by different partners and after thorough evaluation and benchmarking implemented at different access sites (Agilent, Roche/Nimblegen). Exome sequencing using Roche/Nimblegen technology was done in the EOBII Transnational Access Project.

Technology for CNV analysis was improved mainly on the side of using sequencing-based approaches and by the development and implementation of new software. The PeSV-Fisher software was developed (Escaramis et al. 2013 PLoS One) and applied in the Spanish ICGC project of chronic lymphocytic leukemia, which was published recently (Puente et al. 2015 Nature). Another improvement that was heavily used is the no-PCR protocol. This protocol provides very even sequencing depth across the entire GC-range. It turns out that even coverage of a genome is essential for reliable genotype calling (Buchhalter et al. 2014 BioRxiv).

Towards the end of the ESGI project demands for sequencing using input materials extracted from formalin-fixed paraffin-embedded biopsies and specimens increased. In general the fixation process causes extensive damage to DNA and RNA, however, it is possible to extract some viable material. We launched an effort to combined sample preparation for DNA and RNA sequencing starting with DNA and RNA extracted from FFPE material. We found that in particular exome enrichment is compatible with FFPE extracted DNA and results are only marginally inferior to using DNA extracted from a good source.

Several of the developments of this work package have already have impact in that the products were applied across several different studies. The work and experience on eQTLs was used in a large-scale study published in several manuscripts (Lappalainen et al. Nature 2013, t'Hoen et al. Nature Biotechnology). Publications that follow on from this work also benefitted (e.g. Rivas et al. Science 2015). The aforementioned dataset that was significantly funded by ESGI has been placed in the public domain and is accessible and downloadable. The dataset contains roughly 4 Tb of sequencing data. Interestingly it has been downloaded more than 2000 times, which means that this dataset is quite likely used by more than 2000 researchers for further analysis and jointly occupies more than 8 Pb of disc space on computers around the work (for reference the generation of the dataset in total has cost around 300k€,

while researchers deem it acceptable to pay roughly 400€ each for harddiscs to hold their copy of this interesting dataset).

Our work on the development of the no-PCR sample preparation protocol for sequencing has led to us being able to produce that best dataset for the benchmarking of sequencing and mutation calling for the International Cancer Genome Consortium (ICGC, Alioto et al. BioRxiv 2014 and Buchhalter et al. BioRxiv 2014: these two manuscripts are currently under revision for publication in Nature Communication). Further, the improvements of alignment technology have helped us in quality control of the ICGC benchmarks. With the GEM mapper (Marco-Sola et al. Nature Methods, we have developed one of the most performing and accurate aligners that is integrated in the CNAG production pipelines and has been used for the CNAG access projects).

The experience we gained with working on challenging input DNA and RNA has been of great value as more and more projects are reaching back to FFPE resources as this is the only remaining available material. Having worked through this impacts many studies of the future and enables unlocking sample resources that are stored in the past in many hospitals to reach backwards in time.

WP10 (3rd generation sequencing)

A main objective of our RTD activities in this workpackage consisted in the improvement of existing methods for 2nd generation sequencing, i.e. i) replacing older generation of “GA” sequencers from Illumina by the “HiSeq” series of Illumina sequencers (Figure 5); ii) early implementation of commercially available 3rd generation and iii) application of single-cell sequencing technologies. All this work became essential since sensitive sequencing of low abundant DNA fragments or sequencing the entire genomes with 2nd generation DNA sequencers, which were available at the starting point of the ESGI project, was still labour-, time- and cost-intensive.

One highlight was the installation of the substantially evaluated mature HiSeq X Ten system at initially two ESGI sites (WTSI and UU), which allows now for routinely sequencing a human genome with 30x coverage for 1000 Euro within a few hours. Furthermore, alternative sequencing methods such as from Pacific Biosciences (PacBio), Ion Torrent, and early versions of Oxford Nanopore Technologies (ONT) have been tested and applied by ESGI partners.



Figure 5: High-throughput sequencing installation of the 2nd generation at the WT Sanger Institute of ESGI.

Owing to the implementation of new sequencing technologies, ESGI contributed to a number of large-scale genomics projects, for example to decipher the complex aetiology of lymphoblastic leukaemia and to provide potentially useful guidance for targeted therapies (Fischer et al., Nature Genet. 2015).

In the course of the ESGI project single-cell genomics emerged as major issue in the life sciences (Figure 6.). Notably, ESGI already published key articles dealing with method development of single-cell sequencing (e.g. parallel sequencing of single-cell genomes and transcriptomes, see: Macaulay et al., Nature Methods. 2015). Based on refinements made in this WP10 a pilot transnational access project dealing with single-cell sequencing of macrophages exposed to various stress could be launched. Some further functional analyses including single cell analyses of proteins and metabolites will follow after the ESGI project. These experimental steps (which are in part beyond the scope of ESGI) will be required to complete the study to gain fundamental mechanistic insights in cellular resilience to respond to environmental stress.

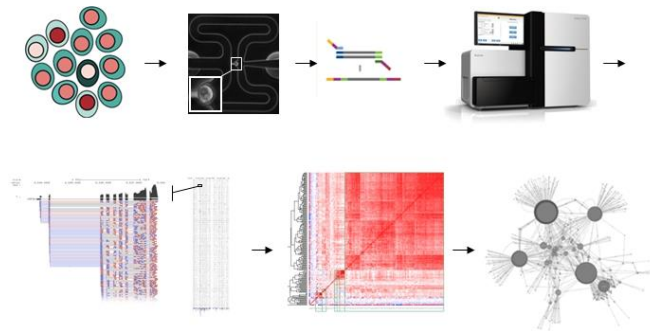


Figure 6: Single cell transcriptome sequencing developed at MPIMG to decipher the role of cell populations in response to environmental changes and in disease processes. Workflow starts with isolation of single cells by microfluidic devices, followed by sequencing using Illumina technology. Sequence data from single cells are used for multivariate and machine learning algorithms to identify cellular subpopulations and gene modules. Further statistical analyses are performed to analyse cellular resilience.

WP11 (Development of data analysis tools)

WP11 focused on the development, implementation, and distribution of bioinformatics and statistical genetics methodologies efficiently analysing sequence variation data.

The methodologies developed for specific applications were made available both as freely available software and as a computer and analysis infrastructure. In conjunction with WP9, a focus was on development of computational tools to develop methods for allele-specific gene expression analysis to map genetic variants that could be functional by regulating the expression and on software for analysing sequence data for miRNA variability. Part of this work has already been published by ESGI participants (e.g. by Lappalainen et al. Nature 2013; Adoue et al. Mol Syst Biol. 2014)

Furthermore, software for fast and reliable mapping of sequencing data has been established to provide an efficient alternative of this key step in almost every genomic application (see e.g. ESGI publication of Marco-Sola et al. Nat Methods. 2012).

When the ESGI project proposed was submitted in 2009, the 1000 Genomes Project was in progress and their agenda included the development of population genetics analysis tools for the vast sizes of genome data. The ESGI project therefore did not aim to duplicate the tool development work within the 1000 Genomes Project. This work package within the ESGI project instead focused on reviewing existing tools and methods, implementing these tools into relevant practical research questions, and developing missing or complementary data analysis methods or tools. The work package had a strong focus on analyzing human clinical biology questions. These include in particular the accurate determination, validation and clinical interpretation of genetic variants, the analysis of gene expression, micro RNA and epigenetic modifications.

The work package was structured into four tasks, and each task was structured into deliverables that individual partners were responsible for. The four tasks in this ESGI work package were:

1. Development of ancillary analytical tools for genetic variant analysis
2. Method for validation of variations (single base changes, small indels, large indels, rearrangements, inversions)
3. Sequence analysis and assignment
4. Statistical analysis

Description of the main results/foregrounds

Under ESGI, a total of 8 methodologies were either extended or newly developed for specific applications and published or made freely available: ASE allele-specific gene expression analysis, ClinCNV detection method for copy number variants in whole exome and targeted next generation sequencing data, GEM accurate and fast alignment software for small computers, miRDeep2 miRNA analysis software tool, pibase SNV validation software, PeSV-Fisher structural variant analysis software, SGA *de novo* assembly tool for small computers, Vy-PER virus integration detection method. To additionally allow analysing dynamics of the genome, we implemented PhantomPeakQualTools (partner CEA-CNG) to analyse transcriptional regulation using ChIP-sequencing (ChIP-seq) and evaluated and fine-tuned methods such as FAIRE-seq, DNaseI-seq and ATAC-seq for analysing transcriptional active or inactive genomic sites.

Bioinformatics toolkits for transnational access projects were installed, and the most generic of these are published on the website: <http://www.sanger.ac.uk/resources/software/esgi/>
A total of 9 peer-reviewed papers were so far published within this work package:

Elsharawy A, Forster M *et al.* Improving mapping and SNP-calling performance in multiplexed targeted next-generation sequencing. **BMC Genomics**. 2012 Aug 22;13:417. doi: 10.1186/1471-2164-13-417.

Forster M *et al.* From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software. **Nucleic Acids Res**. 2013 Jan 7;41(1):e16. doi: 10.1093/nar/gks836. Epub 2012 Sep 10.

Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. **Nat Methods**. 2012 Dec;9(12):1185-8. doi: 10.1038/nmeth.2221. Epub 2012 Oct 28.

Almlöf JC *et al.* Powerful identification of cis-regulatory SNPs in human primary monocytes using allele-specific gene expression. **PLoS One**. 2012;7(12):e52260. doi: 10.1371/journal.pone.0052260. Epub 2012 Dec 26.

Escaramís G, Tornador C, Bassaganyas L, Rabionet R, Tubio JM, Martínez-Fundichely A, Cáceres M, Gut M, Ossowski S, Estivill X. PeSV-Fisher: identification of somatic and non-somatic structural variants using next generation sequencing data. **PLoS One**. 2013 May 21;8(5):e63377. doi: 10.1371/journal.pone.0063377. Print 2013.

Friedländer MR *et al.* Evidence for the biogenesis of more than 1,000 novel human microRNAs. **Genome Biol**. 2014 Apr 7;15(4):R57. doi: 10.1186/gb-2014-15-4-r57.

Petersen BS *et al.* Whole genome and exome sequencing of monozygotic twins discordant for Crohn's disease. **BMC Genomics**. 2014 Jul 5;15:564. doi: 10.1186/1471-2164-15-564.

Forster M *et al.* Vy-PER: eliminating false positive detection of virus integration events in next generation sequencing data. **Sci Rep**. 2015 Jul 13;5:11534. doi: 10.1038/srep11534.

Fischer U, Forster M *et al.* Genomics and drug profiling of fatal TCF3-HLF-positive acute lymphoblastic leukemia identifies recurrent mutation patterns and therapeutic options. **Nat Genet**. 2015 Sep;47(9):1020-9. doi: 10.1038/ng.3362. Epub 2015 Jul 27.






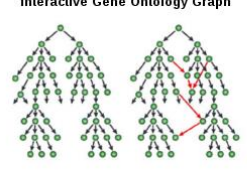
The freely available methods and tools that have been implemented, tested or developed in this work package have produced significant impact as they contributed to increased understanding of high throughput data analysis. Concretely, these methods and tools have been used repeatedly in access-projects or other research projects.

For example, the GEM publication in Nature Methods has been cited 131 times over the last two years, indicating a large user base of the software. The PeSV-Fisher software has already been chosen as one of the two standard tools for structural variant analysis in the American NIH's PanCancer project. The 1000 novel miRNAs found by Friedländer and colleagues are a major breakthrough in miRNA research, and the allele specific expression analysis technique reported by Almlöf and colleagues can dramatically lower the cost of such experiments.

The progress achieved now paves the way for rapid clinical research and next generation sequencing based clinical diagnostic projects. One highlight application utilizing ESGI software tools was the international TCF3-HLF childhood leukemia sequencing project that resulted in a number of new therapeutic options for this fatal leukemia subtype.

In the future, tools for determining the pathogenicity of genetic variation will become accurately estimate the impact on disease. Such tools, co-developed by ESGI partners, are now available for the large user community (Figure 7).

ClinVar & Functional Annotation Tool **cnag**

<p>Clinical Variants</p> 	<p>KEGG Analysis</p> 	<p>Reactome Annotation</p> 	<p>Related literature</p> 
<p>Gene Ontology Table Results</p> 		<p>Interactive Gene Ontology Graph</p> 	

Contacts: eserra@pcb.ub.cat - sbeltrana@pcb.ub.cat

Figure 7: Pathogenicity prediction analysis tools on the access users main results page at Partner 7 (PCB).

1.4 The potential impact

Socio-economic impact and the wider societal implications of the project

Contribution to Community and social objectives

Strategic Impact

The successful establishment of the European Sequencing and Genotyping Infrastructure (ESGI) for providing sustainable support to European researchers was an initiative from 2011-2015 to strengthen the European Research Area (ERA) in life sciences. As the European Strategy Forum for Research Infrastructures (ESFRI) pointed out in the past, the development of an efficient infrastructure for genome research was of crucial importance to position Europe as one of the world-leading regions for genetics, genomics and systems biology research. The technological platforms required for such research are necessary components for maintaining and improving Europe's role in international collaborations with American, Asian and other genomics centres. The ESGI infrastructure could - through its transnational access model - optimise European research programs and foster international collaborations, as well as establish uniformly working conditions for European researchers.

The ESGI initiative contributed to defragment and thereby strengthen European research capacities in genetics and genomics in general, for example by improving knowledge transfer from large genomics centres among themselves and to external expert groups focusing on specific research questions. Thereby, ESGI contributed to a more balanced territorial development of European genetics and genome research.

ESGI could build a momentum to incite the important national genomic centres in Europe to participate in a collaborative effort. ESGI could thus provide a unique frame for producing synergetic effects through integrating complementary expertise from different partners of ESGI to add value to the different institutes by harmonisation of research technologies and methodologies. Consequently, ESGI partners shared expertise and technologies for a larger purpose, and provided benefits for all the partners involved and particularly for external users from small specialised research groups. This particular aim could be achieved by making excellent standardised wet-lab and data handling and data storage facilities accessible to users who would otherwise not have such promising research opportunities.

Technology advances

In a joint effort, ESGI partners agreed on standard protocols and operating procedures covering sample handling, preparation and quality control as well as interpretation and quality assessment of sequencing and genotyping data. This effort contributes to improved reproducibility and reliability of data for future research. Furthermore, by optimising protocols ESGI facilities achieved an overall reduction in operating costs across European centres due to the integrated infrastructure activities. In particular consumable costs for sequencing could be decreased by these efforts. The joint research activities further led to improvements of existing methodologies and resulted in novel methods to solve specific research questions.

ESGI achieved in particular a major impact owing to the introduction of 3rd generation sequencing technologies, by providing routine methods for sequencing the complete human genome within few hours for less than 1000 Euro.

Integration of genetics data from major sequencing/genotyping centres across Europe

This major goal of ESGI was achieved in close collaboration with the ELIXIR project on biological information management through installation of seamless mechanisms for depositing genetic data from the sequencing/genotyping infrastructure into databases such as the European Genome-phenome Archive (EGA). These sequence data resources will provide substantial added value for future analyses. Moreover, data generated by ESGI in the frame of transnational genetic variation projects will serve as European population controls, thus potentially decreasing at least for some scenarios the need for additional costly control genotyping in subsequent projects.

Knowledge transfer

Close interaction of ESGI partners and external scientists, in particular via "hands-on" transnational access projects, fostered cross-disciplinary knowledge exchange by sharing of scientific and technological expertise. Eventually, the provision of world-class high-throughput sequencing and genotyping facilities enabled external European users to generate knowledge rapidly and efficiently, which would otherwise have taken significantly more time and would have required substantial additional funding.

By offering transnational access and extensive training, education of several hundred scientists from a wide variety of institutions across the European Union was achieved by the ESGI project. Knowledge of next generation sequencing and related bioinformatics applications has been disseminated, in particular to junior scientists that will form the next generation of researchers. The training activities of ESGI have coincided with the increasing application of new sequencing methods in the clinical medicine. This in particular shall produce a long-term benefit to public health within the EU.

Furthermore since the market for next generation sequencing technologies and services generates totally tens of billions of Euros per annum, this is bound to produce profound economic benefits right across Europe.

Improved interactions with other European and international initiatives

ESGI has collaborated with multiple international initiatives in the fields of genome research, for the 1000 Genomes and follow-up projects and the international cancer genomics projects (e.g. in the frame of International Cancer Genome Consortium (ICGC)). The applied large-scale sequencing approaches are also being extended to study other diseases and to pursue more diverse systems biology approaches.

To name two outstanding collaborations, ESGI successfully established standards for RNA sequencing in tight collaboration with the European gEUVADIS project (<http://www.geuvadis.org>), as evidenced by two key publications ('t Hoen et al. Nat Biotechnol. 2013; Lappalainen et al. Nature. 2013), and secondly developed sequence data resources such as the European Genome-phenome archive by collaborating with the currently establishing European infrastructure for biological information ELIXIR (<https://www.elixir-europe.org/>) (key publication: Lappalainen et al. Nat Genet. 2015).

To further provide a solid basis for future genomics initiatives, ESGI established collaboration with other existing European infrastructures and resources. Importantly, the ESGI set standards to analyse samples stored in biobanks by collaborating with partners from the Biomolecular Resources Research Infrastructure (BBMRI) and related European projects.

Improvement of European genome research

A number of users, European scientists from many biological and biomedical disciplines, benefited greatly from using modern large-scale genome research platforms offered by ESGI. The efficient installation of the ESGI, which was achieved during its pilot phase, will substantially contribute to allow increasing the number of future projects of external users in an advanced infrastructure.

In general, given the excellent status of its health-care systems, Europe has the capacity to play an important role in the integration of sequence data with related phenotypic information, to improve the understanding of disease and advance diagnostic and therapeutic development. The genetic/genomic information that were generated in the transnational access projects dealing with population and genotype-disease association will contribute to further use the rich phenotypic information of the European sample cohorts. Thus, ESGI contributed significantly to build the ground of a deep knowledge of the genetic variability of Europeans. This shall greatly benefit European biomedical research for the years to come. The storage of all the information at EMBL-EBI databases (coordinated by the partner infrastructure ELIXIR) will guarantee access to these datasets by all investigators. This action will also expand the benefits of the results of ESGI to non-European countries.

Socio-economic issues

European genetics and functional genome research, supported by infrastructures such as ESGI, will increase scientific knowledge to advance the understanding of complex biological processes, as well as the treatment and the prevention of human diseases, among many other biological applications (e.g. biodiversity research).

Transnational access projects of ESGI contributed to a fundamental understanding of the effects of genetic variation in European populations and of basic (gene-) regulation processes involved in disease. This scientific knowledge provides the ground for rationally pursuing alternative targets for medical intervention and for developing powerful diagnostic markers. Qualitatively new knowledge, as gained by the ESGI project, will potentially lead to more efficient treatment strategies in the future. External users and ESGI partners achieved major scientific goals to achieve the here explored overarching goals (as e.g. documented by already published and multiple further high-impact studies to be published within the next years).

As ESGI became a fully operational infrastructure covering key activities of the life sciences and biomedical research, future efforts for establishing a sustainable infrastructure to continuously support the European Research Area are well prepared. Eventually, scientific efforts supported by essential genome analysis infrastructures will contribute to transfer and implement biomedical knowledge into daily practice to eventually improve public health and

decrease financial burdens. This process will come along with knowledge transfer and establishment of the growing European biotechnology and pharmaceutical sector. In the long run research activities such as ESGI shall continue to contribute to socio-economic developments by establishing a rationale scientific basis for an improving living standard in Europe.

Importantly, ESGI did not only focus on technological aspects of the infrastructure, but additionally worked on ethical, legal, and societal issues related to sequence-data derived from humans. Thereby, ESGI interacted closely together with other key infrastructures involved in biobanking (e.g. BBMR) and management of biological information (e.g. ELIXIR), as well as many other stakeholders and the lay public. As described above in more detail, these efforts of ESGI and partner infrastructures will certainly contribute to reasonable implementation of novel sequence analysis technologies, to adequately balance scientific advance, medical progress and the individual interests of the patients.

Main dissemination activities and exploitation of results

Dissemination and Knowledge Management

Dissemination of ESGI focussed on two elements: the dissemination to professionals (both inside the project and the wider community) and the dissemination to the public. Dissemination activities to reach potential users of ESGI are outlined in the workpackage description above (WP6). Furthermore, dissemination of know-how and novel research tools to the scientific community were essential for rapid progress in technology and data analysis tools development and to provide access to state-of-the-art installations (WP4). Publication of methods and data in peer-reviewed scientific journals was thus an important component for dissemination of results to professionals. Thereby, the ESGI adhered to international standards of release of genome data (WP3). ESGI partners will continue to promote a policy of sharing of anonymised data by several means, for example, through the promulgation of a common data archive co-developed by ESGI (WP3). As the feasibility of data sharing may depend upon consent at the point of sample collection, the policy in this regard was carefully considered. Further dissemination channels to professionals were amongst others collaborations with consortia such as gEUVADIS, ELIXIR and BBMRI, multiple presentations of ESGI members at scientific conferences, and participation at various workshops.

The scientific dissemination to integrate sequencing and genotyping technologies and genome analysis across Europe further included for example:

- A regularly updated project website, outlining goals and methodology, providing protocols and online training as well as analysis software, and presenting results.
- Submit internal reports that also provided the basis for the periodic reporting to the Commission.

In parallel ESGI undertook public dissemination. Following the human genome project and the relatively smooth integration of forensic genetics into society, the public's interest in genome and biomedical research has increased significantly in Europe. Public dissemination activities involved for example:

- Public lectures on genetics by investigators of ESGI (such as the Royal Society Summer Exhibition in 2013).
- Contribution to articles in journals and magazines. Interviews with the media.

Exploitation

The exploitation of the results generated by ESGI comprised two main aspects: academic and scientific exploitation. Intellectual property issues played a minor role in the ESGI project, since most of the potentially generated IP lay in the hands of the external users, and because methods and tools implemented by ESGI partners were primarily developed for public free-of charge open-access usage.

The academic exploitation was achieved through training of personnel of ESGI and even more importantly of external users of the European research community. Sharing of methods for sequencing and genotyping, as well as for methodologies in data handling and analyses, proved to be an efficient way to additionally strengthen the research within ESGI, and to significantly advance research capabilities of the external users.

The main scientific exploitation of ESGI consisted in developing and subsequently providing state-of-the-art wet-lab and data analysis tools for transnational access projects. A number of scientific studies are already or will be published in the near future based on the results of the users and ESGI partners. Notably, for further use the data produced in transnational access projects (e.g. data from scientifically selected European populations) were made publicly available (according to ELSI guidelines developed by ESGI) via the European Bioinformatics Institute.

Outlook and future research

We strongly believe that ESGI was a successful pilot project to significantly defragment the European infrastructure capacity in the field of nucleic acids analysis, in particular to advance applications using highly parallel sequencing and related data analysis. It is common sense that this field of research will continue to grow over the next years. To respond to this global trend and the fact that not every small- to medium-scaled research institute would be able to build up the infrastructure required to perform internationally competitive genomics research, ESGI laid the ground work for an advanced distributed infrastructure to allow multiple researchers performing demanding large-scale experiments.

ESGI successfully implemented the framework required for a sustainable infrastructure in multiple ways; technologically, in terms of data production and management, and administratively. Future efforts shall continue to keep the infrastructure at the technological forefront and enable even more European researchers to access high-class facilities, based on the organisational frame developed by ESGI.

A key component of this effort concerns further technology development and early implementation of emerging methods at major European core facilities. Additionally, given the exploding amount of data produced by highly parallel sequencing, intensified development of data analysis tools and interaction with European hubs for managing large-scale biological information such as ELIXIR will be required to adequately handle data of users for future application. As in the past, these continuous developments shall include cooperation with partners and consortia representing standardised biobanks, as well as experts in the fields of ethical, legal and societal issues to ensure proper management of patient data.

Furthermore, based on the now established organisational structure ESGI is a position to significantly increase the number of projects for transnational access in the future. This goal can be further achieved by including additionally technologically and scientifically specialised genome research facilities that emerged over the last years in Europe, for example in the new member states.

We believe that the realisation of the here summarised endeavours to provide and further develop an internationally competitive, sustainable genome research infrastructure would be highly welcome by the large European life science community.