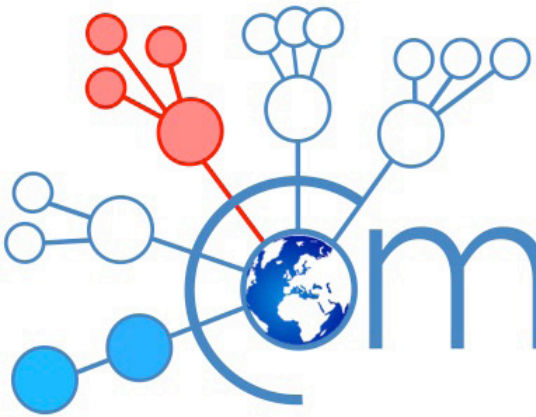


Update - YEAR 1

COMPARE - Year 1
On track and on the right track

During 2015, we have observed a number of larger as well as minor international or global infectious disease emergence and transmission events. This has included ebola, zika and influenza, as well as the newly emerged mcr-1 gene conferring colistin resistance. In all cases, sharing and analysis of sequence-based information in combination with relevant epidemiological information have helped to elucidate the events, but even faster sharing and common analysis between all competing groups would likely have provided even better foundations for public health interventions. Thus, the system and data-sharing platform COMPARE seeks to deliver is more relevant and more needed than ever.

The need for common data-sharing infrastructures are also being increasingly recognised by international organisations, including WHO, FAO, eCDC and EFSA, and the COMPARE consortium has provided input to discussions and supported strategy development to these organizations.

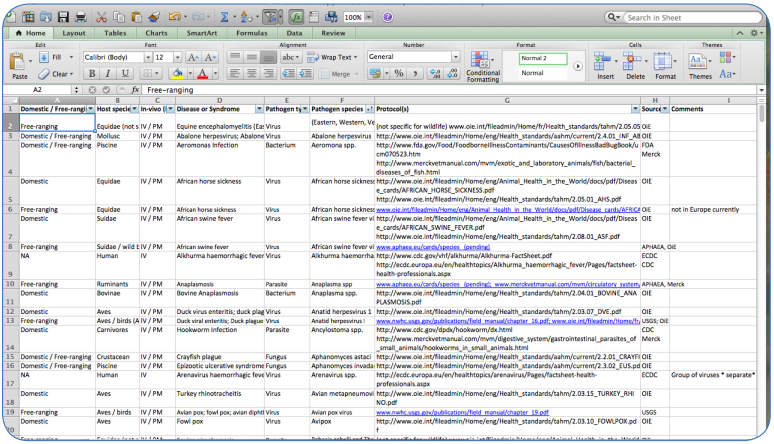
During the first year, COMPARE has developed initial standards or initiated studies for sampling and handling of samples. Initial workflows for clinical diagnostic, food safety and emerging diseases have been developed and pilot projects planned. Initial versions of web-accessible sites for sharing of sequence data have been created, and the first attempts to compare analytic pipelines are on their way. The recent global emergencies have been a good basis for studying communication strategies and plans for economic evaluations will follow.

The activities within COMPARE are on the right track. Many are, as expected at this stage, scattered within individual workpackages, and during the next phase of the project, it is important to integrate the different pieces into a coherent work-flow, which will not only be useful but essential for identifying, tracking and preventing infectious diseases.

WP1

Give me everything you've got – an inventory of sampling protocols

Understanding the extent of existing sampling protocols for generating clinical and diagnostic data arising from food, human, livestock and wildlife populations helps to predict the characteristics of samples that are likely to be supplied or made available through existing surveillance systems. We have been developing an inventory of existing, and where possible, harmonised protocols in order to map the types of samples that are currently recommended at the EU or international level for known diseases of public and veterinary health importance. This inventory is now available for comment by the COMPARE Consortium. The research required was across the different disciplines of human clinical data and the equivalent animal information. This was accessed through the various websites of multi-national organisations and EU-FP7 projects and gave rise to some surprising comparisons. Firstly, accessing livestock information was, overall, easier and structured more logically than datasets held for human health. However, for all areas, information was dispersed and sometimes incomplete, leading to a feeling of a treasure hunt! The accessibility and coverage of datasets made available from several organisations was commended including the World Organisation from Animal Health (OIE) and US Centres for Disease Control and Prevention (CDC). Whilst the inventories are still being completed they have already proven useful for one of the contributors currently working at FAO.



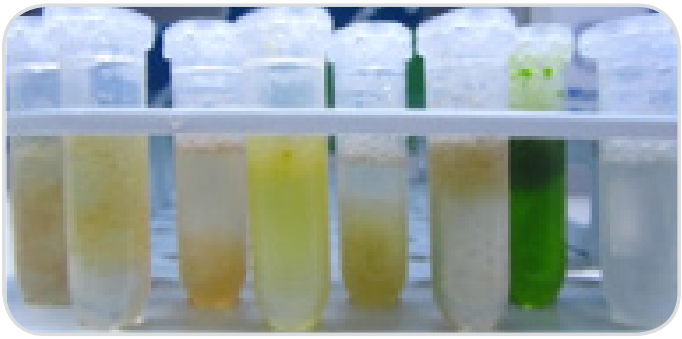
A screenshot from the inventory of protocols

WP2 Everyone has a recipe – optimizing and harmonizing handling protocols

Careful sample handling is a crucial step in gaining high-quality information from Next Generation Sequencing to ensure maximum benefit for clinical and public health. Samples have to be treated with great care to minimize significant shifts in microbial community composition of samples during transportation. This is an important prerequisite in order to display the initial sample situation within the sequencing outcome and to successfully detect causative agents via sequencing. Therefore, work package 2 (WP2) is addressing the harmonization of standards for sample handling as a basis for other tasks in the COMPARE project. During the first year of the project, an inventory of commonly used protocols with respect to collection, handling, transport and storage of samples was conducted via a survey. Based on survey results, experiments were designed to investigate the influence of diverse treatments and handling procedures such as fixation, storage temperature and duration on different sample matrices such as tissue, body fluids, feces, sewage samples, as well as ticks and insects containing pathogens.

In parallel, sample-processing pipelines including pathogen inactivation, nucleic acid extraction, and subsequent processing until sequencing were developed and are being intensively tested for different matrices (e.g., tissue, ticks, bacterial suspensions, food samples). Protocols providing best results regarding quality and quantity of extracted nucleic acids as well as sequence reads were already disseminated in the form of Laboratory Operating Procedures (LOPs) for their further review and application in laboratories of COMPARE members.

These sample-handling experiments and sample-processing pipelines will be tested and further refined, respectively, during the next months. The enthusiastic aim of these optimization steps is to develop and to provide one protocol for all samples for metagenomics.



Different food samples after homogenization using the CryoPrep technique.

WP3 WP6

Going with the flow – from raw data to actionable information, documenting the NGS workflow

The unrestrained use of antibiotics constitutes a major health threat due to acquisition of antibiotic resistance (AR) and rapid spread of multi-drug-resistant bacterial pathogens. In the past few years, several efforts have been made to unify experimental data for easy accessibility for researchers in the form of AR gene data resources. We reviewed the currently available AR gene data resources with the aim of making them more visible to the microbiology research community (Xavier et al., JCM, 2015). Additionally, we also carried out test runs on several databases using in-house and publicly available data. This exercise highlighted the limitations of some of the popular AR gene data resources, with emphasis on the need for regular updates and easy accessibility to resources including metadata from published literature. Finally, we also propose an NGS (metagenomic and whole genome sequencing) pipeline for systematic screening of AR gene resources utilizing the AR gene data resources (Figure 1).

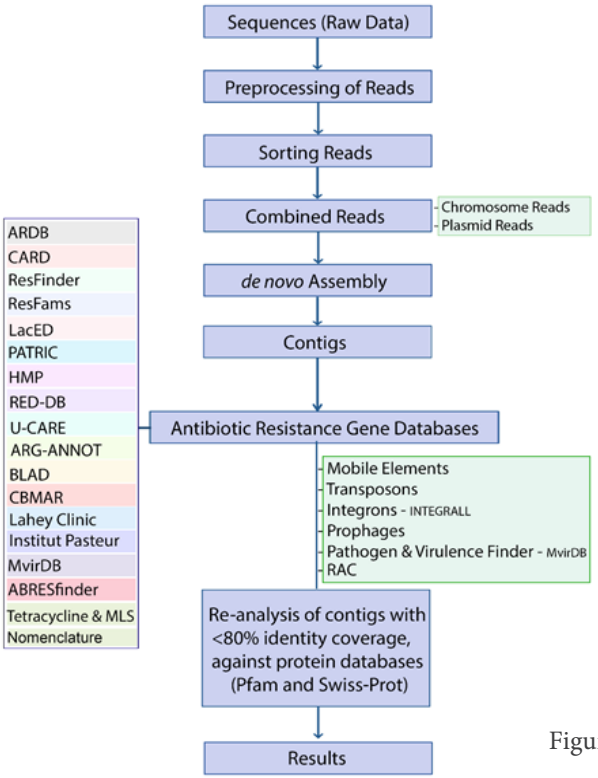


Figure 1

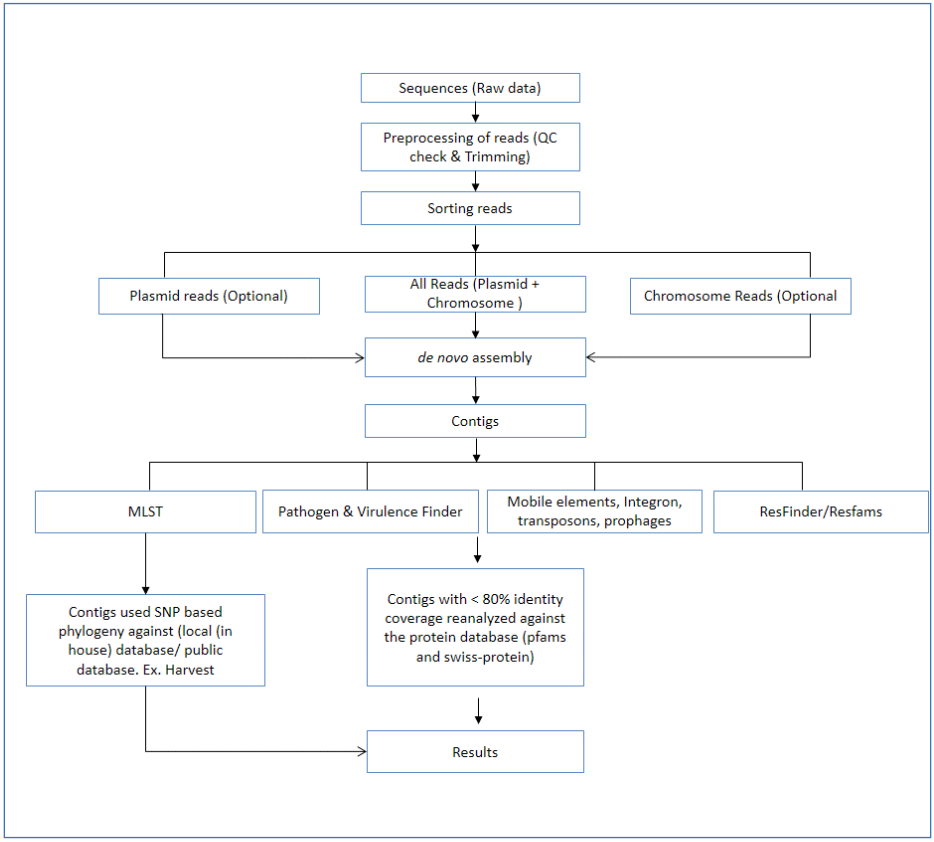


Figure 2. From raw data to actionable information, documenting the NGS workflow

WP4 WP7

Reference Genomes – posting a collection

WP4 and WP7 are focused on developing and validating cross-sector and cross-pathogen methods for sequence-based analysis within surveillance, outbreak investigation, epidemiological analysis, and source attribution of foodborne pathogens.

As a first important achievement in WP4/7, a database of reference genomes covering some of the most important foodborne pathogens of public health relevance was constructed. Specifically, the database included Norovirus, Hepatitis A virus, *Salmonella, Listeria monocytogenes*, verotoxin-producing *Escherichia coli* and *Cryptosporidium*.

The reference genome database was set up in order to provide the first step in standardized, comparable genomic analysis within each of the selected organisms. For each of the six pathogens, the reference genome database was set up to represent the most important types or phylogenetic groups, allowing for reliable and systematic sequence analysis of these microorganisms in the future.

The database was built from publically available genomes. The COMPARE reference genome database, at present, consists of 138 reference genomes (see table) representing 35 bacterial genomes (*Salmonella, Listeria monocytogenes*, and *Escherichia coli*), 88 viral genomes, or partial genomes (Norovirus, Hepatitis A), and 15 genomes of parasitic origin (*Cryptosporidium*).

Pathogen	Number of reference genomes
Salmonella	7
Escherichia coli	10
Listeria monocytogenes	18
Norovirus	79 (complete or partial)
Hepatitis A virus	9 (complete or partial)
Cryptosporidium	15

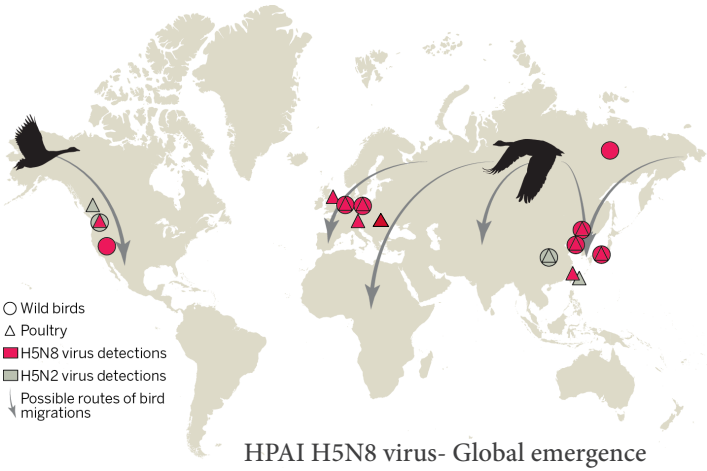
This set of reference genomes is deposited on the European Nucleotide Archive (ENA) web page along with a general background on each of the organisms and their nomenclature. In addition, the rationale behind the specific selection of reference genomes is described for each pathogen.

WP5 WP8

An early start - outbreaks of highly pathogenic avian influenza A/H5N8 virus

In November 2014, just weeks before the start date of COMPARE, highly pathogenic avian influenza A/H5N8 viruses that were previously confined to Asian countries spread to Europe and North America, causing numerous outbreaks in poultry. National and international authorities asked the usual questions: Where did this virus come from, and how did it end up in our backyard? How widespread is this virus in wild and domestic animal populations? How much of a threat is it to animals and humans? At the launch of COMPARE, it was immediately clear that this consortium should collectively employ the available toolboxes to investigate these outbreaks, and evaluate the new tools – which are still under development – in a real-life scenario. These efforts were integral to EU preparedness, gathering scientific evidence underpinning risk pathway identification and better protection of susceptible farmed poultry. Using targeted surveillance approaches, A/H5N8 viruses were rapidly detected in migratory wild birds, and a plausible scenario for their entry into Northern Europe was sketched.

Field and experimental data on infections of avian and mammalian species provided crucial information about potential risks. COMPARE partners brought together a global consortium of investigators with access to virus genetic data from around the world to apply a "forensics" approach in order to trace back the evolutionary and epidemiological history of the viruses causing these outbreaks. COMPARE partners started to compare various pipelines for Next Generation Sequencing (NGS) to investigate the added value of NGS over conventional sequencing methods and potential limitations of the new technology. With the most crucial investigations completed rapidly to inform stakeholders in real time, and other investigations still ongoing to advance technology development in support of future outbreak investigations, the COMPARE consortium clearly demonstrated its ability to respond rapidly to new outbreaks, to deliver critical information on time for risk assessments, while at the same time employing such threats to advance the state-of-the-art of outbreak investigations.



HPAI H5N8 virus- Global emergence

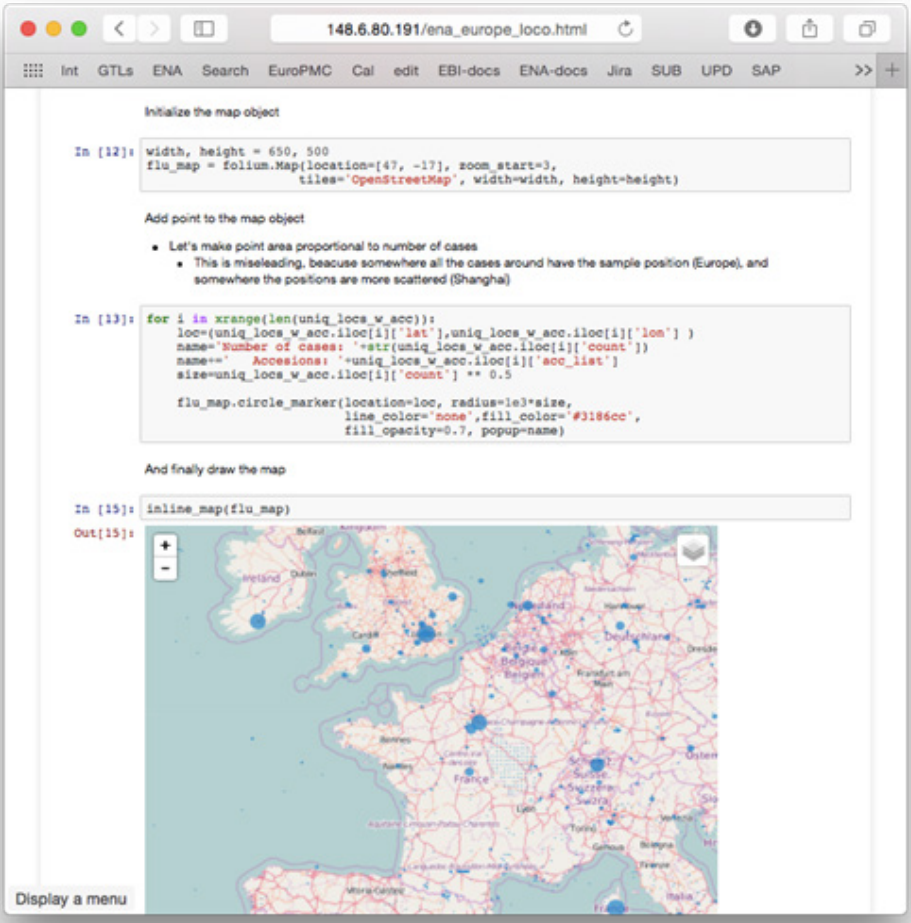
Verhagen et al. Science 347:616-617, 2015

WP9 Sharing and COMPARing

Rapid sharing and analysis of pathogen genomics data is central to COMPARE, and in WP9, we build the informatics to enable these.

Rapid sharing: Two collaborators can easily share and discuss small data files, but add more collaborators and introduce the massive files typically used in genomics and this kind of informal system soon fails. In COMPARE, we follow the principle of 'structured data sharing', in which those providing data for the use of others organise and describe their data systematically in a data reporting step, covering details of sequenced isolates, sequencing methods and file types. Well-structured from the start, these data can be made systematically searchable and will travel with appropriate information to maximise their usability. We have launched the COMPARE Data Hubs that are allowing partners to share data rapidly in this structured way, using a confidential 'quarantine' period prior to full public release.

Enabling analysis: Running analysis on big genomics data sets requires computing power and data to be brought together. We have launched the COMPARE-VM, a cloud compute environment physically co-located with the data storage system that lies underneath the Data Hubs. Providing a host of bioinformatics tools, this is being used by software engineers in COMPARE for the development of computational analysis workflows to be deployed across COMPARE data. For more ad hoc analysis tasks, a 'Notebook' environment is being developed and is already being trialed by COMPARE bioinformaticians.



Part of a Notebook being used for geographical analysis

WP10

Communicating risks can be risky

In the first year, COMPARE WP10 carried out several activities, most of them focused on the identification of relevant stakeholders directly or indirectly involved in risk communication. The work has been developed along three main axes:

1) Identification of communication needs and perception of the issue of risk communication among internal COMPARE stakeholders, i.e., people involved directly in the consortium. This has been explored through personal interactions and a specific survey delivered online to which 53 people responded.

2) Identification of primary COMPARE stakeholders, i.e., individuals and groups who interact, are expected to interact, or could potentially interact directly with the COMPARE consortium. This was performed through an audit of scientific literature, monitoring of official policy documents and reports, and overall monitoring of the institutional links and activities of COMPARE partners.

3) Identification of secondary COMPARE stakeholders, i.e., individuals and groups who will be likely to interact with primary stakeholders, and thus indirectly also with the COMPARE consortium. These are the "final users" of COMPARE outcomes; they include the larger stakeholder communities. This task has been performed by auditing the relevant scientific literature, scanning official policy documents and reports, and regularly monitoring the social conversation in more than 900 web sites, blogs and Facebook pages, and in a specifically devoted Twitter list including approximately 500 people. This material has been further processed by using tools for content and network analysis (Nvivo 11 plus), which allowed the extraction of themes, relevant people, and influencers.

Risk communication is not really important, the real issue is science communication	Risk communication is the same in any field	Risk communication in public health is different from risk communication in any other field	Risks related to infectious diseases have specific communicational needs	Communicating the risk of emerging pathogens with outbreak potential needs a sp. approach	Total	Weighted Average
2,33%	0,00%	4,65%	37,21%	55,81%		
1	0	2	16	24	43	4,44

Figure 1: 93.2% of COMPARE participants think that risks related to infectious diseases have specific communication needs.

WP12

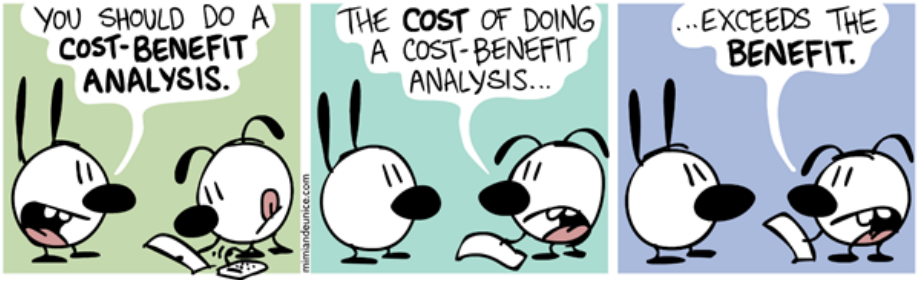
Constructing a One Health pathogen database in Europe – building on uneven terrain

Creating one NGS database for a wide variety of pathogens on a European, and in the end, global scale, leads to many technical, language and bioinformatic issues. On top of these, other obstacles of different classes and nature are to be encountered. In line with more general focused public health studies[i], these are identified as motivational, economic, political, regulatory, legal and ethical barriers.[ii] We note that stakeholders, both scientists as well as health authorities, are often left in uncertainty about the interpretation and consequences of complex national and international legal frameworks. This easily leads to a conservative not-sharing or not timely sharing reaction. Clarification of international conventions such as WHO International Health Regulations, the Nagoya Protocol on genomic resources or EU regulations on data protection, cross border health threats, and dual-use of biological materials and information, is of utmost importance, as a starting point for solutions for real existing barriers. On the one hand, the magnitude of problems may well appear larger than they actually are.[iii] This has to be sorted out. On the other hand, remains the fact that COMPARE brings together different worlds and domains, with economic interests and legal frameworks that are divergent, and in different stages of international harmonization. We are building on uneven terrain. The interaction between barriers to data sharing in public health is complex, and single solutions to single barriers are unlikely to be successful.[iv] A successful construction demands to focus on sustainability and equity.

[i] van Panhuis et al., A systematic review of barriers to data sharing in public health, BMC Public Health 2014, 14:1144 [http://www.biomedcentral.com/1471-2458/14/1144]
[ii] Sane, Jussi and Michael Edelstein, Overcoming Barriers to Data Sharing in Public Health: A Global Perspective, Chatham House publication, 2015, ISBN 978 1 78413 050 3.
[iii] Aarestrup, F.M. and M.G.Koopmans, Sharing Data for Global Infectious Disease Surveillance and Outbreak Detection, Trends in Microbiology, 2015, 1293
[iv] See e.g. first conclusion of Sane e.a. 2015.

WP14 COMPARE – can we expect a lot of bang for the buck?

Although COMPARE is not for free and costs money, COMPARE potentially brings about huge benefits through a variety of mechanisms such as earlier detection of disease outbreaks but also through increased research output. However, quantifying and valuing the benefits is often more challenging than quantifying the costs. WP14 has the aim to quantify costs and benefits of the COMPARE system and will develop methods to value the benefits of COMPARE.



www.compare-europe.eu

Twitter @CompareEurope

COMPARE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 643476.

YEAR 1