Jigsaw Genomics

Assembling the pieces towards open and accessible bioinformatics for everyone

Saskia Hiltemann

Jigsaw Genomics

Assembling the pieces toward open and accessible bioinformatics for everyone

Jigsaw genomics

De puzzelstukjes op hun plaats leggen voor open en toegankelijke bioinformatica voor iedereen

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

TBD

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on

Wednesday 27 October 2021, at 10.30hrs
by

Saskia Desirée Hiltemann
born in Tilburg, The Netherlands

Erasmus University Rotterdam

Erasmus

Doctoral Committee:

Promotors:
prof. dr. P.J. van der Spek
prof. dr. ir. G.W. Jenster

Other members:
prof. dr. S.G.M.A. Pasmans
prof. dr. M.J. Roobol - Bouts
prof. dr. P.-.B. 't Hoen

Copromotor:
dr. A.P. Stubbs

This one is for me.

Propositions accompanying the thesis

Jigsaw Genomics
Assembling the pieces toward open and accessible bioinformatics for everyone

by Saskia Hiltemann.

1. The Galaxy platform increases accessibility of bioinformatics analyses and enhances reproducibility of scientific research (this thesis)
2. High-quality bioinformatics training is an integral part of empowering domain experts to analyze their own data (this thesis)
3. iFUSE and Circos visualizations were instrumental in identifying fusion genes and chromothripsis in the VCaP cell line (this thesis)
4. A large, representative virtual normal set is equally valuable as an associated normal sample for germline correction of tumour samples (this thesis)
5. The MYcrobiota platform improves standardization and reproducibility across experiments, studies, and laboratories for microbiota profiling in diagnostics (this thesis)
6. The COVID-19 pandemic has illustrated the power of, and need for Open Science. (10.1101/2020.08.13.249847)
7. Efforts to attract more women to STEM (Science, Technology, Engineering and Mathematics) should target parents, teachers and the general public, not only young women. (10.1007/s11199-011-9996-2)
8. More effective communication of scientific results to the general public, and in particular the robustness of the conclusions drawn from the results, is instrumental for increasing public trust in science and reducing the spread of misinformation.
9. Academia should reward the creation, and long-term maintenance of software tools as much as they value new publications and awarded grants to combat the short lifespan of bioinformatics tools.
10. When we increase diversity in academia, we all win (10.15252/embr.202051994)
11. Humans are allergic to change. They love to say, "We've always done it this way." I try to fight that. That's why I have a clock on my wall that runs counter-clockwise. - Grace Hopper

o

Foreword

The big red button. It is every biologist's dream; a mythical and mystical red button that takes their raw data and transforms it magically into a *Nature* paper ready for submission. Of course this is an unattainable dream, but there are many steps we can take to ease the burden of data analysis and decrease the turnaround time between sample collection and manuscript submission.

The road towards the big red button is made up of many pieces, which when combined in the right way will server to demystify bioinformatics and reveal the final the final image; a world in which researchers are empowered to once again analyze their own data without needing specialist bioinformatics skills.

Contents

"This job is a great scientific adventure. But it's also a great human adventure. Mankind has made giant steps forward. However, what we know is really very, very little compared to what we still have to know."

Fabiola Gianotti

0 Introduction

0.0 The Source Code of Life

DNA. The blueprint of life. These long double-stranded helical molecules are present in all living cells on earth* and encode the proteins which drive the functioning, structure and replication of the cells and tissues that make up an organism. In many ways, DNA is also analogous to computer code; any computer program, no matter how complex, can be described as a long series of just two characters, `0` and `1`, known as *bits*. Knowing the sequence of these bits and, crucially, the details about how they are being interpreted by the machine on which they are executed, enables us to understand and predict the functions they encode. In much the same way, DNA uses just 4 different elements, called *bases* or *nucleotides*, to encode its blueprint for the cell. These 4 building blocks are adenine, cytosine, guanine and thymine, usually referred to simply by their first letters, `A`, `C`, `G` and `T`. These bases combine together in pairs (*base pairs*), with `A` always matching to `T` and `C` being complementary to `G` along a sugar phosphate backbone in their

*Some viruses contain only RNA, but these are often not considered 'alive'.

double-helix configuration (Figure 0). Using continually improving and increasingly affordable genome sequencing techniques, are now able to read this string of bases that make up the cell's blueprint, but unravelling the mysteries of how these are being interpreted by the cell poses a far greater challenge.

In humans, DNA is organised into 23 pairs of chromosomes, with a total length of approximately 6.4 billion base pairs. Sprinkled across these chromosomes are *genes*; stretches of DNA which often encode protein molecules. When these protein-coding genes are expressed, the gene sequence is transcribed to a messenger molecule (*mRNA*) which is subsequently translated into a protein molecule. About 1% of the total DNA consists of *exons*, the portions of genes that directly encode protein sequences. The function of the remaining 99% of the genome long remained a mystery, and was even referred to as *junk DNA*. More recent studies have revealed that many of these stretches of non-coding DNA play an important role in the regulation of the expression of genes, stimulating or prohibiting the translation of genes to proteins and thereby influencing the functioning of the cell. For any two individuals, the vast majority of their DNA sequence will be identical, but small variations in the remaining locations of the genome are what make each of us unique. However, if these variations occur in the wrong place, it also has the potential to cause illness. Studying this natural variation in the genome sequence helps us unravel the mechanisms of life and disease.

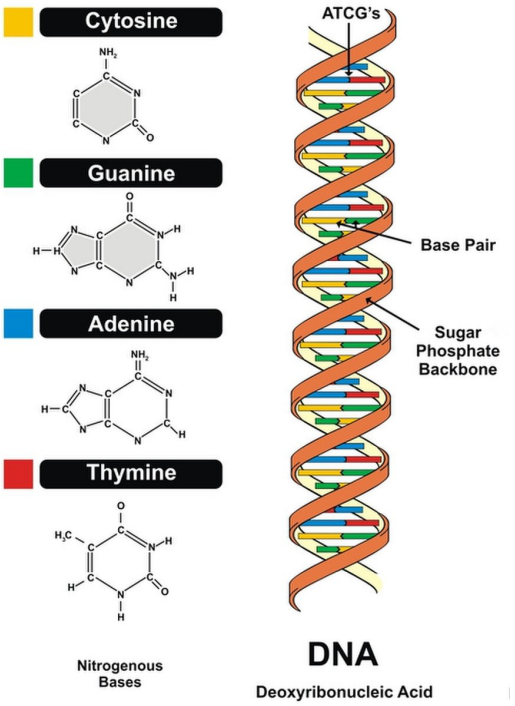


Figure 0: Structure of DNA

A Brief History of Genome Sequencing

Friedrich Miescher was the first to isolate DNA molecules (which he termed *nuclein*) in 1869. However, the molecule remained relatively understudied until Rosalind Franklin used X-ray crystallography to inspect the molecule's structure [0]. Watson and Crick famously used this data to postulate the double-helix model of DNA in 1953 [1]. This model illustrated for the first time that DNA molecules were ideally suited for replication, and solidified the idea that DNA, and not protein as was previously thought, was the primary carrier of hereditary information. Simultaneously, Frederick Sanger had developed techniques to sequence proteins, and later RNA and DNA molecules. This was the start of the *first generation* of genome sequencing, and the first full gene was sequenced in 1972, followed by the first full genome in 1976 (bacteriophage ϕX174, 5386 nucleotides long) [2]. These early techniques were only suitable for relatively small sections of DNA, and more technological advances would be required before larger genomes could be fully sequenced with greater coverage.

In 1990, the Human Genome Project [3] set out to sequence the entire 3.2-billion-basepair-long human genome, an effort culminating in 2003 with the publication of the first human *reference genome* [4]. Not only did this provide invaluable insights into human genetics, but it also paved the way for the next era in genetic research; something which would completely transform the field of genetic research. Over the next several years, *next-generation* massively parallel sequencers were developed by companies like Roche454, Solexa, Complete Genomics, Illumina, and others, dramatically cutting the cost and time required to sequence a human genome, and for the first time demonstrated its potential utility in clinical and diagnostic settings. Through sustained technological advancements over the following years, these costs continued to decrease at exponential rates - outstripping even the pace predicted by Moore's law (Figure 1) - and the long dreamed-about *$1,000 dollar genome* [5, 6] has now become a reality.

Shortly after the publication of the human genome, high-throughput RNA-seq techniques emerged to sequence the *mRNA* molecules that are transcribed from the DNA, collectively termed the *transcriptome*. This allowed for the identification and quantification of gene transcripts, providing valuable information about which genes are *expressed* (active) at a given point in time, and at which levels. Furthermore, this also has the capability of providing researchers with information about post-transcriptional mutations and other complexities such as alternative splicing [8]. Similarly, the study of epigenetics is revealing that non-coding DNA, far from the once-termed *'junk DNA'*, is part of an intricate and complex regulatory system controlling the expression of genes [9].

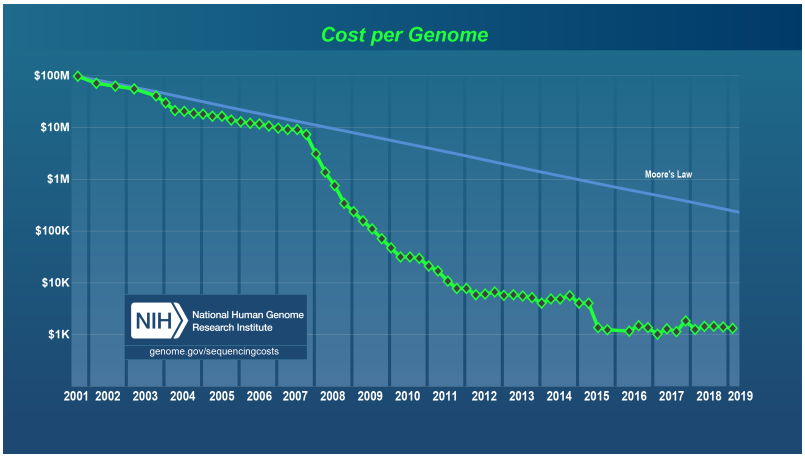


Figure 1: The cost of sequencing a human-sized genome over time. Data from the NHGRI Genome Sequencing Program (GSP) [7].

Now we are moving into the era of *third-generation* sequencing, where single-cell [10], long-read [11], and often real-time sequencing [12] are allowing for ever more accurate determination of nucleotide sequence, providing increased resolution even in highly diverse and complex samples such as cancer or metagenomic samples. All these technological advances have led to a deluge of data that must be managed and analysed, typically by bioinformaticians.

0.1 The Bioinformatics Challenge

With huge amounts of data now being generated at relatively low cost [13], and compute resources being available even on moderate budgets, the challenge in genomic research has shifted from the sequencing technologies to the analysis and interpretation of these big and highly complex datasets, and the development of the software required in order to gain new understanding of the underlying biological systems. This poses significant challenges, both technologically and scientifically.

A jigsaw genome

Current sequencing techniques cannot simply read all the nucleotides of an entire chromosome at once; instead, DNA molecules are cut into tiny pieces of a few hundred bases, and each of those small fragments is then sequenced. The full genome must then be reconstructed from these short sequence reads. Think of this as solving a jigsaw puzzle with billions of pieces. To complicate matters further, the nucleotides determined by the sequencer (*base calls*) are not always correct. Moreover, we typically don't sequence the DNA from just one cell, but combine material from a

large number of cells together and leading to a lot of duplicate and overlapping pieces. And then of course, the genome itself contains complexities such as repetitive regions, that complicate the reconstruction further. Luckily however, in the case of human DNA, we have the picture on the puzzle box to guide us, in the form of the reference genome. This reference genome can be used as a scaffold to which to map each of the reads, improving our chances of successfully reconstructing the genome from the short read data. However, a reference genome does not exist for all organisms, and diseases such as cancer can incur high degrees of genomic rearrangements, decreasing the utitlity of the reference genome. In such cases, the short reads must be assembled *de-novo* in order to reconstruct the genome.

Ever-changing landscape

Sequencing technologies are evolving at a staggering pace, and with them, so must the software tools capable of analysing the data. Therefore, new tools are developed continually and existing tools must be updated regularly to remain relevant. As a consequence, typically there simply is not enough time for community consensus and data and analysis standards to emerge organically before the advent of new technologies make them obsolete. Because of this, there usually are a multitude of tools available for any given task, each with their own set of advantages and disadvantages, and some of the main challenges are finding the right tool for the particular situation, and knowing when to use existing tools and when the development of novel tools is in order.

Standardisation

A typical bioinformatics analysis will consist of many different tools, chained together into a *pipeline*. In order for data to pass smoothly from one tool to the next, it is important to agree on standard file formats. Unfortunately, in the absence of an authoritative body to prescribe such standards, these do not exist in many cases, and where they do, they often allow for a large degree of flexibility. As a result, files cannot always be passed seamlessly from one tool in the pipeline to the next, and custom file transformation steps are often required as a *glue* of sorts between the different analysis steps.

Data Storage

It has been predicted that genomics may soon overtake other *Big Data* domains such as astronomy and video-hosting platforms such as Youtube in terms of yearly data generation [14]. Handling these vast amounts of data is a challenge many bioinformaticians face. Not only must the data be stored somewhere where it can be easily accessed, it must also be organised, backed up, and shared

in an efficient way. Any storage solution must be able to scale to the exponential trend of growth both in size and number of files. Furthermore, data storage solutions must comply with privacy and security requirements, which is especially important in the case of human genetic data.

The Specialist Bioinformatician

All of these challenges have resulted in bioinformatics becoming a highly specialized field, which in itself poses a new challenge given the observation that the domain knowledge (biology) and the informatics know-how more and more often do not reside in a single individual, and the interpretation of the data cannot always be done by the person performing the data analysis and vice versa [15]. Instead, close communication between the two fields is required, and ideally the domain experts should be empowered to perform their own day-to-day data analyses without the need of a bioinformatician.

0.2 Bioinformatics Best Practices

In order to optimally deal with all the challenges in bioinformatics, adhering to a set of best-practices guidelines may be beneficial. Such best practices have been described many times [16, 17, 18, 19, 20, 21, 22], with the most notable example being the widely-adopted FAIR (Findable, Accessible, Interoperable, Reusable) data principles [23]. The FAIR data project provides guidelines and tools to improve standardisation within bioinformatics.

Findability

The first step in (re)using scientific datasets, tools, or services, is to find them. Simply making these resources available online is not enough to make them truly findable. By annotating datasets and tools with metadata using standardized ontologies (e.g. EDAM Ontology [24]) using standardised syntax schemes (e.g. BioSchemas [25]), allows them to be discovered more easily by both humans and computers. This will also allow services to be created to aggregate this information into discovery portals for users (e.g. BioTools [26] for analysis tools and databases, and the TeSS portal [27] for training materials).

Accessibility

Accessibility of tools and data implies not only that these resources be made publicly available (e.g. on GitHub for code, or Zenodo for data), but also that they are easily findable and *usable*. Most bioinformatics tools are command line UNIX [28] tools, and biologists are not typically trained in the use of such. Even for the experienced bioinformatician, running some of these tools can be a

challenge due to lack of documentation or quality of the tool. Ideally, once a tool or pipeline has been validated, the analysis can be run by the domain expert, i.e. the research scientist responsible for the interpretation of the analysis results, without being reliant on the support of a bioinformatician at every step [29].

Creating user-friendly software is not a trivial task. Application linking (also referred to as *wrapping*), can ease this burden for the tool developer. In such an approach, existing user-friendly interfaces host third-party software packages –at minimum effort to the developer of the hosted software– and thereby offer a layer of abstraction to the end-user that shields them from the implementation details of the tool and provides a uniform usage paradigm for all tools, regardless of their differences behind the scenes. Two examples of such hosting frameworks in the context of bioinformatics are Galaxy [30, 31] and Taverna [32].

Accessibility also includes the creation of high-quality documentation of tools, both aimed at developers and end-users, and ideally some form of training manual to educate users in the proper use of tool and to warn about possible pitfalls and biases. Platforms like GitHub are not not only great for version control of the code, but also allow for integration of documentation that lives with the code, and integrations with 3rd-party

Reproducibility

A cornerstone of the scientific method is reproducibility of results. Experiments should be described in sufficient detail to allow for their reproduction and independent verification by fellow scientists. In theory, since bioinformatics occurs entirely in the digital realm –and is not dependent on physical samples as is in the wet lab– it should be possible to completely reproduce an experiment provided that all the input data is available and the analysis is sufficiently well described. In reality however, this remains a big challenge, and many publications cannot be reliably reproduced based on the information provided therein [33, 34].

As a result, reproducing bioinformatics analyses often becomes an exercise in *forensic bioinformatics*; attempting to piece together the exact procedure used through trial-and-error and educated guessing, rather than by the information provided by the original researchers. This is not without consequences; in some cases this has lead to clinical trials being started based on incorrect conclusions not revealed during peer review due to lack of reproducibility [33].

In recent years, many scientific journals have started taking steps towards remedying this situation, by improving submission guidelines for bioinformatics pipeline descriptions, similar to those they impose for the physical samples. The latter are commonly required to be submitted to biobanks

before submission, and the raw datasets generated from them to be stored in online repositories. For true reproducibility of scientific findings, a standardised description of the full software stack used to analyse the data, as well as a detailed description of how this code was applied to the raw data in order to arrive at the presented conclusions must be provided.

One obstacle to the reliable replication of bioinformatic studies arises from the fact that, while reproducibility is a high priority in the scientific community, this is not true for software developers in general, and bioinformatics pipelines are often dependent on a combination of scientific and more general-purpose components. For example, many software packages will simply be removed once an update has been made available, hindering scientific reproducibility of all but the most recent analyses. This is where package managers such as Conda [35] or GNU Guix [36] and containner technologies such as Docker [37] offer significant improvement, by allowing any historic versions of tools to be installed at all times [38]. This aims to ensure that the full stack of software and dependencies obtained during installation of the pipeline will yield identical results when performed today as it will a year or 5 years from now.

However, in addition to describing the complete software stack used for analysis, the pipeline *provenance* must also be captured; metadata describing *how* the tools were applied to the data in the analysis run. This includes the sequence and interplay of different tool executions, with the full set of parameters, and the input and reference data used at every step. Keeping a lab journal of the in-silico experiment is a good start, but is too error-prone as a manual process. Projects such as Jupyter Notebooks [39] for Python, or Sweave [40] and KnitR [41] for R, allow for the intermixing of text and code to create a kind of interactive journal article. Calculations can be embedded within the text, and these calculations may be examined and rerun or adjusted with ease by the reader. A drawback of this approach comes when tools require a lot of resources or are not all written in the same language, as is typically the case for bioinformatics analyses. Workflow platforms such as Galaxy automatically keep track of provenance for the user and have the advantage of supporting a wide range of tools and programming languages.

Interoperability

The aforementioned lack of data format standards hinders the interoperability of tools. Even in cases where file formats are somewhat standardized, variations in the exact implementations may still require careful consideration within a pipeline. Consider for example the FASTQ format; this is a widely used and relatively simple format, it consists of 4 lines per sequence read, but the line containing quality score, is the source of some divergence in the format. Different tools and sequencing platforms use different encoding schemes, but this information is not stored within the

file itself, making it hard to ascertain which encoding scheme was used without proper metadata accompanying the file. Knowing the quality encoding scheme used for any datasets entering your pipeline is of critical importance, as low-quality reads could be mistaken for high-quality reads and vice versa if incorrect assumptions are made. Similar widespread variations in standard formats include chromosomal location (0-based or 1-based numbering, open or closed) and chromosome names (with or without a `chr` prefix) or gene names (many different naming schemes). These issues are not hard to deal with on a conceptual level, but in practise can easily lead to inaccurate results if not carefully taken into account by the creator of the pipeline.

Such concerns exist on the more purely informatics side of bioinformatics as well. Take the example of command line shell scripts, many OS-dependent syntax variations can hinder interoperability, but by taking care to comply with POSIX standards [42], such concerns can be mitigated.

Going beyond file formats, interoperability of different tools may be hindered due to the fact that they are often written in different programming languages and/or designed for different operating systems. Fortunately, tool developers can undertake steps to make their tools more interoperable. Package managers such as Bioconda [43] will compile software from its source for a variety of different operating systems and simplify installation for end users.

Finally, and vitally, precise documentation of any assumptions made in the code is indispensable, and making source code open further facilitates this transparency and interoperability [44].

Maintainability

Since scientific research depends increasingly on software, decreasing the burden of tool maintenance is valuable and worthwhile pursuit. Furthermore, many tools are written by small research groups or single individuals who are simply unable to perform the necessary maintenance without support. By making tools open-source, the entire bioinformatics community is able to step up as co-developers, allowing them to discover bugs and contribute fixes or enhancements. Code sharing platforms like GitHub [45] and BitBucket [46] facilitate this community-driven approach to software maintenance.

Using code versioning such frameworks such as git [47] or mercurial [48] ensures precise tracking of changes and enable collaborative tool development. Incorporating tests at every phase of development can further decrease the maintenance effort. Unit tests ensure that small code modules show expected behaviour at all times. Functional tests ensure that these different units of code always yield the desired result when working in conjunction. Code quality checks can help

streamline the code itself and increase readability, benefiting future development. Continuous integration is the paradigm whereby changes to the mainline code are incorporated incrementally and continually, and thoroughly reviewed and tested at every stage, thereby decreasing the maintenance burden of the software. Code sharing platform such as GitHub and BitBucket offer continuous integration frameworks as part of their service.

0.3 Bioinformatics for Everybody

In order to make bioinformatics accessible to a wider audience, and empower domain experts to analyze their own data, several components are required. Firstly, analyses must be made available in a user-friendly way, not requiring any expertise in programming or systems administration to install and run the pipelines. Secondly, data visualisation and integration are essential components to aid the interpretation of results by the research scientist in this *big data* era. Finally, extensive training of researchers is crucial in order to provide them with the required knowledge and confidence to run analyses and interpret outcomes independently.

The Galaxy Project

The Galaxy project [49] provides a user-friendly graphical interface to command line tools, bringing the data analysis to the domain experts equipped to perform the interpretation of the results. It also facilitates many of the best-practice bioinformatics guidelines outlined in the previous section; with its heavy focus on accessibility and reproducibility, Galaxy is a potent framework for creating high-quality bioinformatics pipelines. Galaxy keeps track of the full analysis provenance, manages tool dependencies with Bioconda [43], has built-in visualisations, and is accessible by end users with nothing more than a web browser. For developers, Galaxy provides a convenient framework to package tools to make these available for researchers throughout the global community. The Galaxy tool shed [50] is the central repository of tools that anybody may contribute to, and contains over 7200 tools at the time of this writing. As such, it is a powerful resource for making bioinformatics accessible and reproducible.

Visualisation and Reporting

As scientists become increasingly reliant on large and complex computational analyses in their research [13], the final analysis result datasets become similarly complex and have often grown beyond the realm of what can be manually viewed and interpreted, both in terms of the number of files and their sizes, as in terms of complexity. Analysis results therefore require summation

and visualisation and must be presented to the domain expert in a comprehensible and accessible manner. Such a report should also contain detailed descriptions of the methods used and assumptions made and citations to any third-party tools employed, as an understanding of these factor can assist in interpretation [29].

For genomics data, visualisation tools such as Circos [51] and MultiQC [52] enable the integration of various output datasets, often in the order of millions of lines each, to be summarized in a single image or report. While some resolution may be lost in such visualisations, they do enable the easy identification of areas of interest and guide the interpretation by pointing the domain experts in the direction of further inspection.

Training

With -omics research becoming increasingly computational in nature, and many research groups not having access to enough dedicated bioinformaticians, there is a great need for high-quality bioinformatics training to ensure that the domain experts who interpret the results of data analyses are optimally equipped to do so. Surveys confirm this need for bioinformatic training; the majority of researchers (>95%) work with or plan to work with large datasets, but most (>65%) possess only minimal bioinformatics skills and are not comfortable with statistical analyses [53, 54]. Demand for training currently greatly exceeds the supply [55]. In a recent survey [56] over 60% of biologists expressed a need for more training while only 5% called for more computing power. This indicates that the true bottleneck of the current data deluge is not storage or processing power, but rather the knowledge and skills to utilize existing resources.

With its focus on accessibility and user-friendliness, the Galaxy platform is not only a great workflow management system, but also an ideal environment for teaching. Trainees are shielded from the minutiae of the implementation details of the underlying tools, and need nothing more than a browser to execute the tutorials. Because Galaxy can host any command line tools and the tools available in the tool shed cover a wide range of topics, creation and maintenance of a set of training materials must be a collaborative community effort, with content being created by a large number of people with expertise in the various topics.

Bioinformatics in the Clinic

With the decreasing costs of next-generation sequencing, and the increased accessibility of the tools and compute resources required for analysis, bioinformatics is increasingly being incorporated directly into the clinic. In contrast to research applications, clinical bioinformatics directly informs

disease management and patient care. Therefore, careful validation of all tools and pipelines is required to ensure quality of the results. Furthermore, there is great need for standardisation of pipelines to ensure a uniform standard of care across clinics. Finally, there is a greater need for data security and privacy as compared to research applications, as clinical data usually involves patient identifiable data. To this end, a set of 17 best-practice recommendations for the validation of clinical NGS-based bioinformatics pipelines have recently been published by the Association of Molecular Pathology (with support from the College of American Pathologists and the American Medical Informatics Association) [57]. While these recommendations focus on variant analysis, most guidelines are applicable more broadly. These guidelines include practical recommendations for the design, development, and operation of the analysis pipelines, with an additional emphasis on the importance of properly trained and qualified molecular professionals. These clinical guidelines recommend that laboratories perform their own validation, led by a medical professional well-trained in NGS interpretation. The validation process should closely mimic the real-world scenarios faced by the lab. Validation should occur both on each component of the bioinformatics pipeline individually, as well as the full end-to-end functionality of the pipeline as a whole. Additionally, the set of QC metrics employed to determine the performance of the pipeline must be evaluated as part of the validation process. Furthermore, since there is a general lack of standardization in bioinformatics, clinical validation must specify and ensure standardization of nomenclature throughout the pipeline. This entire process of validation must be repeated whenever any component in the pipeline is altered. Furthermore, validation must ensure that pipelines are compliant with any applicable lab accreditation standards, as well as comply with the relevant laws concerning data security with regards to patient identifiable data. Moreover, clinicians must be alerted to any unintended or unauthorized changes to data files.

Many of these recommendations have significant overlap with general bioinformatics best-practice guidelines. For example, bioinformatics best-practices call for extensive code review and unit- and functional testing, as well as continuous-integration testing, standardisation of file formats and nomenclature, and data security and integrity checks. By adhering to these bioinformatics guidelines from day one of development of any tools and pipelines, we ease the burden of clinical validation, and increase the likelihood of adoption of tools and pipelines within the clinic.

Data protection

In clinical bioinformatics, data are typically pseudo-anonymised before analysis, decoupling the sequencing data from the clinical metadata, and assigning an anonymous identifier for the sample. This allows the sequencing data to be analyzed externally, while the clinical data resides in separate data repositories such as OpenClinica [58], only accessible to a limited number of clinicians directly

involved in the patient's care. However, as sequencing techniques become more accurate, and our knowledge of the relationship between genotype and phenotype increases, privacy concerns also grow, and we must be continually re-evaluate the security risks and their mitigations when dealing with NGS data [59].

Use Cases

The following sections describe the two real-world research areas where bioinformatics is of great importance; cancer genomics and clinical microbiota profiling. In both examples, implementation of the bioinformatics concepts and best practices described above provide valuable benefits to the research scientists.

The Jigsaw Genome

I like puzzles. Any type of puzzle. I always have. If I see a puzzle or a problem I have to solve it. I think that is what makes cancer such a fascinating topic for me.

Imagine you are given a jigsaw puzzle. Now instead of a few hundred pieces, there are several billion pieces. The picture on the box is not a picture of the puzzle inside the box, it is just a somewhat similar image. Oh, and did I mention there are a whole bunch of pieces missing? And that many pieces are duplicated? Some pieces don't even belong in our box, but come from a completely different puzzle. On top of that, your little sister has spilled paint over some of the pieces so those can't be trusted to reliably contribute to the image. And instead of one single puzzle, the box contains several; they are all variations of the image on the box, but you have no idea how many different puzzles the box contains. Sound challenging? This is the problem we are solving whenever we sequence a cancer genome.

[Metaphor key]

puzzle pieces = sequence reads
picture on box = reference genome
missing pieces = hard-to-sequence areas
other puzzles = contamination
painted pieces = sequencing errors
multiple puzzles in box = clonality

0.4 Use case 1: Prostate Cancer

> *The time has come in America when the same kind of concentrated effort that split the atom and took man to the moon should be turned toward conquering this dread disease.*
>
> President Richard Nixon

On December 23, 1971, President Richard Nixon, buoyed by recent technical feats such as the moon landing, signed into law the National Cancer Act, thereby declaring a war on cancer. Today, more than 45 years later, that war is still being waged in full force. While great advancements have been made towards this goal, some of the initial optimism has been quelled by discoveries of the great complexity and heterogeneity underlying cancer.

The Hallmarks of Cancer

Tumor cells evolve from normal cells through the acquisition and accumulation of mutations. The human body has mechanisms in place to repair or dispose of damaged cells and to prevent runaway cell division, so in order for a tumour cell to survive and thrive it needs to acquire changes that provide it with advantages for proliferation and evasion of the cell's defense mechanisms. Evidence suggests this transformation from healthy cells into malignant cells follows a strikingly similar path across all different tumour types [60]. A cell's acquired abilities that drive tumour progression are known as the *hallmarks of cancer* [60, 61, 62, 63, 64] and includes such traits as evasion of immune response and programmed cell death, self-sufficiency in growth signals and insensitivity to anti-growth signals, limitless replicative potential, sustained angiogenesis (creation of new blood vessels), tissue invasion and metastasis, reprogramming of energy metabolism, genome instability, and tumor-promoting inflammation.

Each of these steps overcomes one of the body's anti-cancer defense mechanisms. This evolution into malignancy is an Darwinian process where the mutations acquired are random, but those cells that have gained mutations which are advantageous for survival will be able to replicate and thrive and accumulate further mutations. Distinguishing the mutations that impart a strategic advantage and thereby *drive* a tumour's progression, from the often huge number of less harmful *passenger* mutations accumulated over the lifetime of a cancer cell is crucial to our understanding of cancer progression [65]. The optimal course of treatment for a patient often depends on the mutations present and how the cell functions are subsequently impacted by those mutations.

Cancer's Complexities

Determining the exact genetic sequence of healthy individuals is already quite a challenging endeavor; trying to extend this to cancer genomes takes this challenge to the extreme. There are several complexities present in cancer genomes that make accurate determination of the genetic changes and their downstream impacts a difficult task. In the following sections we will discuss some of these complexities and explore the informatics challenges they pose.

The Bio

Small variants comprise the simplest class of mutations; those consisting of alterations of just a handful of bases, for instance the *substitution*, *deletion* or *insertion* of one or more nucleotides.

The impact of such mutations depends on where in the genome they occur. Single nucleotide variants (SNVs) in exonic regions can range from having no effect on the resulting protein (silent), to changing an amino acid in the protein to a different amino acid (missense mutations), to changing a codon into a stop codon (nonsense mutation), which nearly always results in a nonfunctional protein. If this happens in a protein that is vital to the functioning of the cell this can lead to a range of health conditions [66].

While variants within the coding sequence are most likely to have an impact on cell health, small variants *outside* the coding sequence can also have drastic impact on health [67]. For example, 70% of melanomas exhibit a point mutation in one of two positions in the promoter region of *TERT* [68, 69], suggesting such somatic point mutations in regulatory regions may play a role in tumorigenesis. Therefore, whole-genome sequencing may reveal valuable information not detected using exome sequencing.

Structural Variants (SVs) are larger-scale mutations involving rearrangement of segments of DNA of more than roughly 50 bp. In some cases, these rearrangements can cause (parts of) two genes that are usually separated by some distance on the genome to come together to form new hybrid genes. In some cases, these fusion genes may be transcribed into fusion proteins, with a potentially disruptive effect on cell processes. The text book example of such a gene fusion is the so-called Philadelphia fusion frequently observed in leukemia [70, 71]. In this Philadelphia chromosome, a translocation between chromosomes 9 and 22 leads to a fusion of the *BCR* and *ABL1* genes. This fusion gene is expressed, and the aberrant fusion protein causes disruptions to key signalling pathways governing the cell cycle, causing the cells to divide uncontrollably and thereby drive tumor progression. Accurate detection of such structural variations is crucial, as they may serve as diagnostic markers [72, 73] or even therapeutic targets [74, 75].

SVs most often occur outside coding regions, but this does not diminish their capacity for great impact, for instance by disruption of the regulatory mechanisms of tumour suppressor genes or oncogenes [76]. These types of SVs not involving coding regions directly, may only be detected through whole-genome sequencing using paired-end sequencing. In rare cases, a phenomenon known as *chromothripsis* may occur. In contrast to the more common gradual acquisition of mutations over time, chromothripsis involves a shattering of (part of) a chromosome in a single catastrophic event, and the subsequent imprecise stitching back together of the genetic sequence by the cell's repair mechanisms. This results in a cluster of thousands of SVs in a confined genomic region. Chromothripsis occurs in about approximately 2%-3% of cancer genomes, with a significantly increased rate incidence in some cancer types [77]. Identification and reconstruction of such highly rearranged cancer genomes in order to predict their effects remains an open challenge [78, 79].

Tumor heterogeneity refers to the phenomenon that cells in different physical areas in a tumour may be genetically very distinct. When a tumour cell divides to create two new cells, these daughter cells will acquire new mutations.

Over time, this may lead to the formation of different clusters of cells each with a different set of acquired mutations. As a cell obtains a mutation that is beneficial to proliferation, it may instigate a new cluster, or *clone*, of genetically similar cells while in other areas of the tumour, cells may follow a different evolutionary path and create genetically different sub-clones (Figure 2) [80].

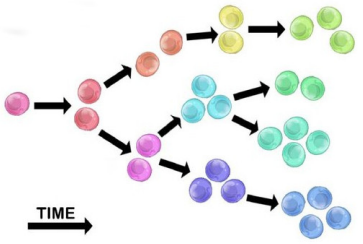


Figure 2: As tumour cells divide, they may acquire additional mutations. Different areas of the tumour may follow different evolutionary paths, leading to tumour heterogeneity.

When sequencing a tumour, DNA from different cells –and thus potentially very different genomes– is mixed together and sequenced as one. Reconstructing the different clonal genomes from this data is immensely challenging.

Further complexities in the analysis of cancer genomes include the temporally evolving nature of tumours; performing the same sequencing experiment on the same tumour at different points in time will often yield a very different picture as the composition of the tumour and

the acquired mutations will have changed. This further complicates the process of comparing different samples and elucidating key characteristics that have the potential to aid in the diagnosis and treatment of patients. Furthermore, studies in epigenetics have shown that tumorigenesis is not solely driven by mutations in nucleotide sequence, but secondary alterations such as DNA methylation may have an equal if not greater impact [81].

The Informatics

All these complexities in the biology of cancer genomes translate to complexities of the downstream informatics analysis pipelines.

Mutations in DNA are described relative to a *reference genome*, and comparing variants across different samples is inherently challenging for a variety of reasons. Observed differences between a sample and the reference genome could indicate a true biological variant, or be the result of a sequencing or mapping error, and separating the true variants from the noise is a task that each variant calling tool approaches differently. As a result, different variant callers often have poor overlap [82, 83]. This poor concordance combined with the observation that there often exist multiple different yet equivalent ways of describing a variant (Figure 3), which hinders the comparison of variants between different samples and studies [84].

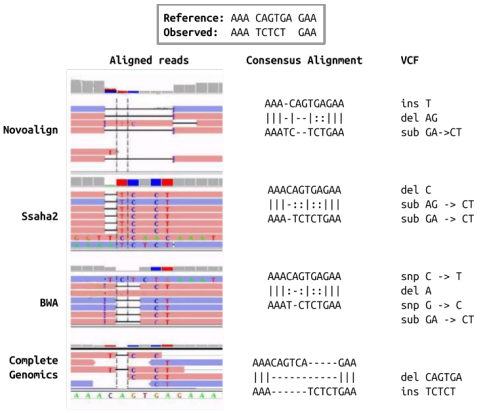


Figure 3: Complex variants can be represented in multiple ways. Four different aligners (Novoalign, Ssaha2, BWA, Complete Genomics) treat the same variant in vastly different ways, which leads to such differing sets of variant descriptions in the VCF files that it is no longer apparent that these variants in fact describe the same observed sequence. Image adapted from the Genome in a Bottle Consortium [84].

In cancer studies often aim to distinguish the *somatic* mutations –those acquired by the cells during tumorigenesis– from the individual's *germline* variants that are present in every cell. To this end, normal tissue is often sequenced alongside the tumour material, and the resulting variant sets

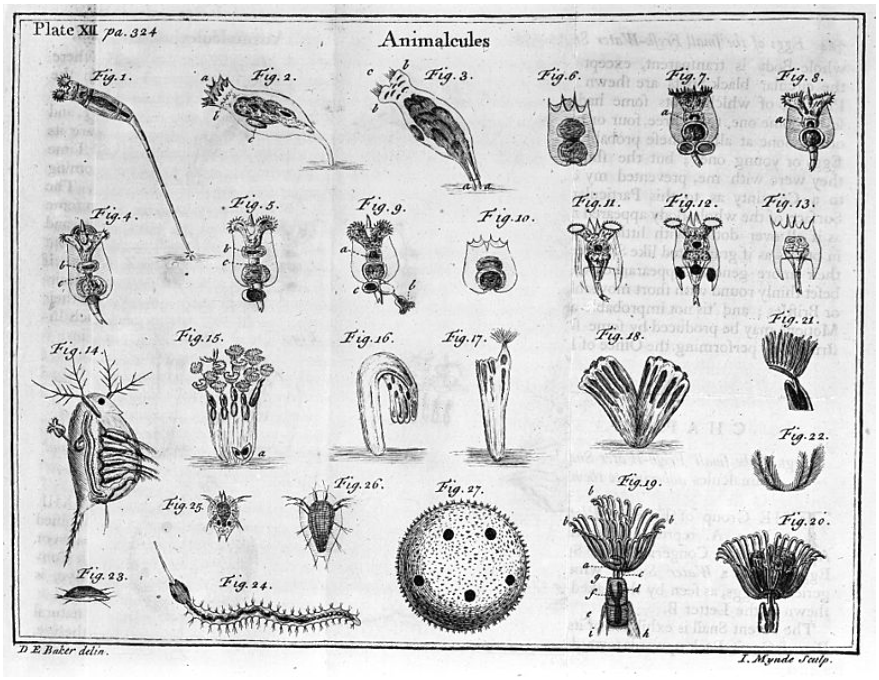
are compared to elucidate the set of mutations only present in the cancerous DNA. Therefore, having accurate variant callers and intelligent methods of variant comparison is of vital importance. However, even methods designed specifically for the identification of somatic variants [85, 86, 87] show poor agreement [88, 89, 82], and improvements are continually sought [90, 91].

Structural variations are by definition large-scale variants, and most sequencers use short reads, making SVs dificult to detect. Consequently, the concordance between different SV calling methods is even poorer than for the small variants [92]. This issue is compounded by a lack of a standard file format [93] making comparisons between studies very difficult. The advent of third generation (long-read) sequencing offers prospects of improving SV detection [94, 95]. These methods typically have higher error rates than short read sequencing techniques, but by using a hybrid approach the advantages of long read sequencing can be combined with the accuracy of short reads to optimize detection of large scale genomics rearrangements [96, 97, 98, 99].

The Human Microbiome

"By the means of Telescopes, there is nothing so far distant but may be represented to our view; and by the help of Microscopes, there is nothing so small as to escape our inquiry; hence there is a new visible World discovered to the understanding. By this means the Heavens are open'd and a vast number of new Stars and new Motions, and new Productions appear in them, to which all the ancient Astronomers were utterly strangers. By this the Earth it self, which lyes so neer to us, under our feet, shews quite a new thing to us, and in every little particle of its matter, we now behold almost as great a variety of Creatures, as we were able before to reckon up in the whole Universe itself."

Robert Hooke, 1665 (in the Preface of Micrographia)



0.5 Use case 2: Microbiota profiling

In 1682, Dutch inventor and scientist Antonie van Leeuwenhoek first turned his microscope to a drop of rain water and discovered within a wealth of microscopic life which he termed *animalcules*. With this finding, a whole new world of knowledge opened up, as it became clear that while these organisms might be microscopic in size, their capability to impact on human lives was enormous, causing food spoilage and disease. Through continued study over the following centuries, links between particular microorganisms and disease were discovered, and remedies developed.

More recently, researchers have realised the importance of not only studying the relationship between specific microorganisms and disease, but also of the structure and composition of the human microbiome *as a whole* with respect to human health. This observation has led to the launch of large-scale research efforts such as the Human Microbiome Project [100] and MetaHIT [101] to investigate the links between the microbiome and health, and kickstarted a wealth of new research.

The Bio

Every cell in our body contains a copy of our genome in its nucleus, which has been dubbed the blueprint of life. But these cells are not the only source of genetic material in our bodies; each one of us harbour a vast amount of micro-organisms at various anatomical locations. Estimates put the number of microbial cells as equal to or outnumbering human cells [102]. These microorganisms influence our metabolism, impact our health and immunity, and may affect drug efficacy. If our genomes are the blueprint of life, then our microbiomes are an important overlay to this blueprint; or, when we consider the genome as the *source code* of life, the microbiome is an essential *third-party plugin*. For this reason, the microbiome is often referred to as our *second genome* [103], and has emerged in recent years as an important field of biomedical study. Links between the microbiome composition and diseases have been demonstrated in a number of diseases [104] including psoriasis [105], obesity [106, 107], and colorectal cancer [108, 109], though in many cases a causal link remains to be proven.

The Informatics

Sequencing the human microbiome comes with its own unique set of challenges. Because the microbiome typically consists of a large number of different organisms, this situation is akin to solving hundreds or thousands of jigsaw puzzles simultaneously, with all the pieces mixed together in a single box. Depending on the research objective, different approaches may be used

to sequence the microbiome. Amplicon sequencing is a targeted approach where a (part of a) single gene is sequenced, which is sufficient for taxonomic identification of the organisms in the sample. In order to obtain additional information about the functional profile of the organisms in the community (for example screening for antibiotic resistance genes), whole-genome *shotgun* sequencing is required [110].

Amplicon sequencing targeting the 16S rRNA gene is used in situations where we are primarily interested in the *composition* of the microbiome; in discovering *which* microorganisms are present, without looking deeper into their exact genomic sequences to ascertain functional information. Consider for example patients with a bacterial infection which, depending on the exact type of bacteria present, may call for treatment with different antibiotics.

With 16S profiling, we do not sequence the entire genomes of the microorganisms, but only a (part of a) single gene of each genome, the 16S rRNA gene. This gene is ideally suited for this purpose because it is present in all bacteria, but contains enough variability between different species to be able to distinguish between them. Because only a small stretch of DNA is sequenced, this approach is much cheaper than were we to sequence all DNA. Think of this approach as looking only at the edge pieces of the jigsaw puzzles; you will not be able to see all the details of the puzzle, but you may have just enough information to deduce whether the picture on the box is a portrait or a landscape.

Whole-genome shotgun sequencing is used in instances where we are interested not only in the composition of the microbiome, but also in its *functional* characteristics. For example, antibiotic resistance of bacteria has become a major concern in recent years; bacteria evolve and obtain mutations that make antibiotics less effective, and identifying the presence of these mutations may influence diagnostics and treatment options. While this approach is more powerful, it is also more costly and the data analysis more complex [110].

Clinical Diagnostics While these NGS-based methods are commonplace in research studies, their uptake in routine clinical diagnostics remains limited. Instead, laboratories employ a host of alternative methods (both culture-based and culture-independent) to determine the cause of microbial infections. However, these methods are not always capable of identifying the infectious agent, as many microorganisms require highly specific growth condition for culturing that cannot easily be replicated in a laboratory setting [111]. Similarly, culture-independent methods such as targeted PCR require *a priori* knowledge of microorganisms suspected to be

present in the sample under investigation. These limitations allow unexpected or hard-to-culture causative agents to evade detection [112]. NGS-based methods certainly possess the potential to deliver such clinically relevant information, by virtue of their culture-independent sample-agnostic nature. However, several obstacles preventing their introduction into routine clinical diagnostics remain, such as issues surrounding standardization, result reproducibility and accessibility. To overcome these obstacles, concerted efforts toward standardization and validation of pipelines are required [113].

0.6 Scope of this Thesis

Bioinformatics has become and integral part of biomedical research, and vast amounts of *-omics* data are being generated. In contrast to other big-data producing fields such as astronomy, biologists are not typically trained in programming and command line usage, which are essential skills for running bioinformatics analyses. The aim of this thesis is to facilitate accessible bioinformatics and to empower researchers in the biomedical field to perform their own day-to-day analyses, without the need to become fully-fledged bioinformaticians themselves. This requires a conscious focus on accessibility, training, bioinformatics best practices, FAIR data principles, and open science in the development of bioinformatics tools and pipelines.

In order to facilitate these goals, some bioinformatics foundations are necessary (Chapters 1–3). This foundation could then be utilized and expanded upon to facilitate a range of scientific research projects. We applied this approach to three such scientific research projects in the fields of tumour variant analysis and microbiome profiling (Chapters 4–6).

Technical foundation

Data analysis platform

In Chapter 1.1 we address the question of how we can make bioinformatics tools and pipelines more accessible for non-expert users. With the Galaxy platform we aim to provide a technical framework to support the full end-to-end flow of scientific analyses, including the efficient management of large datasets, the configuration of tools and workflows, the presentation of results, and the sharing of data and analyses. The results of NGS analyses often comprise large collections of big datasets, which cannot be assessed without an aggregation step. In Chapter 1.2 we were interested in providing rich visualisation options within the Galaxy framework, in order to support the summarization of large genomic datasets and aid in the interpretation of analysis results. Finally, most bioinformatics pipelines consist of a large number of different analysis components, resulting

in a large number of output files. Combining these results into a single coherent and informative view is essential both for research purposes and clinical applications. In Chapter 1.3 we set out to provide a customizable and flexible reporting module for the Galaxy framework that allows end-users to define interactive web reports for their analysis pipelines.

Training

With the technical framework in place, the next step towards improving the accessibility of these bioinformatics analyses comes with providing comprehensive training. Galaxy greatly simplifies the process of running complex data analyses, but a basic understanding of the underlying computational methods is vital for the accurate interpretation of results. Therefore, extensive training materials should accompany these analysis pipelines in order to educate the users on their use, strengths and limitations, and effect on downstream analysis. In Chapter 2 we address the question of how to best facilitate the collaborative creation of bioinformatics training materials aimed at domain specialists in a scalable and sustainable way.

Use-cases

In the remaining chapters, we illustrate this open and accessible approach to bioinformatics through a set of use cases. Each chapter contains a paper about a bioinformatic tool developed, and a paper about the biological study it enabled.

Cancer Analysis

In **Chapter 3** we are interested in the detection and characterization of structural variants in tumour DNA samples, and study their potential to result in fusion genes and proteins. Fusion genes may severely disrupt the cell's processes, and can serve as important biomarkers for diagnosis [114, 115] or therapeutic targets [74, 75]. When sequencing tumour samples, especially those with high degrees of genome rearrangements, a large number of structural variants are typically detected. However, most SVs will not result in fusion genes. Furthermore, chromothripsis is a phenomenon often occurring in in tumours which results in an unusually large number of SVs, and may play a role in tumour progression. In Chapter 3.1 we investigate how best to facilitate the interactive exploration of SVs and identification of potential fusion genes in an user-friendly visual application. We subsequently set out to evaluate the utility of this application in conjunction with tools developed in Chapter 1, by using it to characterize SVs and identify potential fusion genes within the VCaP prostate cancer cell line (Chapter 3.2).

When investigating variants in tumour samples, we are often interested in ascertaining which variants are somatic (acquired during tumour progression), and which are part of the individual's

germline variants. To this end, when DNA from a tumour is sequenced, a second sample is taken from healthy tissue of the patient. Variants present in the normal sample are then subtracted from the tumour variants, in order to find the set of tumour-specific variants. However, in some cases such an associated normal sample is not available (e.g. for public data), or cannot be sequenced (e.g. due to increased sequencing costs or unavailability of the tissue). In these situations, we need an alternative method for discriminating somatic mutations from germline variants. In **Chapter 4** we address the question of whether a large set of healthy unrelated samples (a *virtual normal*) may serve as a viable alternative to a matched normal sample.

Microbiota profiling

Different scientific domains may involve very different analyses, but many of the bioinformatics foundations are shared. Similarly, while the previous chapters focused on research applications, the same bioinformatics approach can be applied to clinical use-cases as well. To illustrate this, in Chapter 5 we investigated the utility of the Galaxy ecosystem for the development of a diagnostic application for use in a clinical setting (MYcrobiota). With sequencing costs decreasing rapidly, NGS analyses are making their way into the routine diagnostics practices. *Streeklab Haarlem* is a medical microbial lab servicing GPs and hospitals across the western part of The Netherlands, providing diagnostics services for (suspected) microbial infections. They employ culture-based methods to determine the cause of the infection, and while this is a tried and true method, it is not always possible to determine the infectious agent using this method due to the fact that some microorganisms require highly specific growth conditions that cannot be efficiently cultured in a lab (e.g. anaerobic bacteria). For these situations, NGS methods such as 16S sequencing may offer a viable alternative, as this approach does not suffer from the same limitations. However, Streeklab Haarlem had no prior experience with bioinformatics data analyses, and reached out to us for a collaboration project. Together we developed and validated the analysis pipeline, and ensured that the platform was accessible for non-bioinformaticians, enabling Streeklab analysts to handle the routine day-to-day data analyses with minimal training.

Bibliography

[0] L. O. Elkin, "Rosalind franklin and the double helix," *Physics Today*, vol. 56, no. 3, pp. 42–48, 2003.

[1] J. D. Watson, F. H. Crick, *et al.*, "Molecular structure of nucleic acids," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[2] F. Sanger, S. Nicklen, and A. R. Coulson, "Dna sequencing with chain-terminating inhibitors," *Proceedings of the national academy of sciences*, vol. 74, no. 12, pp. 5463–5467, 1977.

[3] M. V. Olson, "The human genome project.," *Proceedings of the National Academy of Sciences*, vol. 90, no. 10, pp. 4338–4344, 1993.

[4] I. H. G. S. Consortium *et al.*, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, no. 7011, pp. 931–945, 2004.

[5] "Prize overview: Archon x prize for genomics.." https://web.archive.org/web/20080421165009/http://genomics.xprize.org:80/genomics/archon-x-prize-for-genomics/prize-overview.

[6] "National human genome research institute (nhgri) dna sequencing costs: Data." https://www.genome.gov/sequencingcostsdata/.

[7] W. KA, "Dna sequencing costs: Data from the nhgri genome sequencing program (gsp)." https://www.genome.gov/sequencingcostsdata. Accessed 02-09-2019.

[8] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, p. 57, 2009.

[9] E. Zuckerkandl and G. Cavalli, "Combinatorial epigenetics,"junk dna", and the evolution of complex organisms," *Gene*, vol. 390, no. 1-2, pp. 232–242, 2007.

[10] C. Gawad, W. Koh, and S. R. Quake, "Single-cell genome sequencing: current state of the science," *Nature Reviews Genetics*, vol. 17, no. 3, pp. 175–188, 2016.

[11] S. Koren and A. M. Phillippy, "One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly," *Current opinion in microbiology*, vol. 23, pp. 110–120, 2015.

[12] B. A. Flusberg, D. R. Webster, J. H. Lee, K. J. Travers, E. C. Olivares, T. A. Clark, J. Korlach, and S. W. Turner, "Direct detection of dna methylation during single-molecule, real-time sequencing," *Nature methods*, vol. 7, no. 6, p. 461, 2010.

[13] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile networks and applications*, vol. 19, no. 2, pp. 171–209, 2014.

[14] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big data: Astronomical or genomical?," *PLOS Biology*, vol. 13, p. e1002195, jul 2015.

[15] L. Preeyanon, A. B. Pyrkosz, and C. T. Brown, "Reproducible bioinformatics research for biologists," *Implementing Reproducible Research*, p. 185, 2014.

[16] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig, "Ten simple rules for reproducible computational research," 2013.

[17] T. Seemann, "Ten recommendations for creating usable bioinformatics command line software," *GigaScience*, vol. 2, Nov. 2013.

[18] G. Wilson, D. A. Aruliah, C. T. Brown, N. P. C. Hong, M. Davis, R. T. Guy, S. H. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, *et al.*, "Best practices for scientific computing," *PLoS biology*, vol. 12, no. 1, p. e1001745, 2014.

[19] A. Prlić and J. B. Procter, "Ten simple rules for the open development of scientific software," 2012.

[20] S. Altschul, B. Demchak, R. Durbin, R. Gentleman, M. Krzywinski, H. Li, A. Nekrutenko, J. Robinson, W. Rasband, J. Taylor, *et al.*, "The anatomy of successful computational biology software," *Nature biotechnology*, vol. 31, no. 10, p. 894, 2013.

[21] A. Via, T. Blicher, E. Bongcam-Rudloff, M. D. Brazas, C. Brooksbank, A. Budd, J. De Las Rivas, J. Dreyer, P. L. Fernandes, C. Van Gelder, *et al.*, "Best practices in bioinformatics training for life scientists," *Briefings in bioinformatics*, vol. 14, no. 5, pp. 528–537, 2013.

[22] F. da Veiga Leprevost, V. C. Barbosa, E. L. Francisco, Y. Perez-Riverol, and P. C. Carvalho, "On best practices in the development of bioinformatics software," *Frontiers in Genetics*, vol. 5, July 2014.

[23] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, p. 160018, 2016.

[24] J. Ison, M. Kalaš, I. Jonassen, D. Bolser, M. Uludag, H. McWilliam, J. Malone, R. Lopez, S. Pettifer, and P. Rice, "Edam: an ontology of bioinformatics operations, types of data and identifiers, topics and formats," *Bioinformatics*, vol. 29, no. 10, pp. 1325–1332, 2013.

[25] A. J. Gray, C. A. Goble, R. Jimenez, *et al.*, "Bioschemas: From potato salad to protein annotation.," in *International Semantic Web Conference (Posters, Demos & Industry Tracks)*, Vienna, Austria, 2017.

[26] "Bio.tools - a comprehensive registry of software and databases." https://bio.tools.

[27] "Training esupport system (tess)." https://tess.elixir-europe.org.

[28] "Unix: a family of multitasking, multiuser computer operating systems." https://www.opengroup.org/membership/forums/platform/unix.

[29] S. Kumar and J. Dudley, "Bioinformatics software for biologists in the genomics era," *Bioinformatics*, vol. 23, no. 14, pp. 1713–1717, 2007.

[30] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko, "Galaxy: a platform for interactive large-scale genome analysis," *Genome Res*, vol. 15, no. 10, pp. 1451–1455, 2005.

[31] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome biology*, vol. 11, no. 8, p. R86, 2010.

[32] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li, "Taverna: a tool for the composition and enactment of bioinformatics workflows," *Bioinformatics*, vol. 20, pp. 3045–3054, jun 2004.

[33] K. A. Baggerly and D. A. Berry, "Reproducible research," *Journals must begin to demand*, 2009.

[34] Y.-M. Kim, J.-B. Poline, and G. Dumas, "Experimenting with reproducibility: a case study of robustness in bioinformatics," *GigaScience*, vol. 7, jun 2018.

[35] "Conda - package, dependency and environment management for any language." https://conda.io.

[36] L. Courtès, "Functional package management with guix," *arXiv preprint arXiv:1305.4584*, 2013.

[37] "Docker: Package software into standardized units for development, shipment and deployment." https://www.docker.com/.

[38] N. Kulkarni, L. Alessandrì, R. Panero, M. Arigoni, M. Olivero, G. Ferrero, F. Cordero, M. Beccuti, and R. A. Calogero, "Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines," *BMC bioinformatics*, vol. 19, no. 10, p. 211, 2018.

[39] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay, *et al.*, "Jupyter notebooks-a publishing format for reproducible computational workflows.," in *ELPUB*, pp. 87–90, 2016.

[40] F. Leisch, "Sweave: Dynamic generation of statistical reports using literate data analysis," in *Compstat*, pp. 575–580, Springer, 2002.

[41] Y. Xie, "knitr: a comprehensive tool for reproducible research in r," *Implement Reprod Res*, vol. 1, p. 20, 2014.

[42] S. R. Walli, "The posix family of standards," *StandardView*, vol. 3, no. 1, pp. 11–17, 1995.

[43] B. Grüning, R. Dale, A. Sjödin, J. Rowe, B. A. Chapman, C. H. Tomkins-Tinch, R. Valieris, J. Köster, *et al.*, "Bioconda: a sustainable and comprehensive software distribution for the life sciences," *bioRxiv*, p. 207092, 2017.

[44] D. C. Ince, L. Hatton, and J. Graham-Cumming, "The case for open computer programs," *Nature*, vol. 482, no. 7386, p. 485, 2012.

[45] "Github: a web-based hosting service for version control using git." https://github.com/.

[46] "Bitbucket: a web-based version control repository hosting service." http://www.bitbucket.org.

[47] "Git: a distributed version control system." https://git-scm.com/.

[48] "Mercurial: a distributed revision-control tool for software developers." https://www.mercurial-scm.org/.

[49] E. Afgan, D. Baker, M. Van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, *et al.*, "The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W3–W10, 2016.

[50] "Main galaxy toolshed." http://toolshed.g2.bx.psu.edu/.

[51] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, "Circos: An information aesthetic for comparative genomics," *Genome Res*, vol. 19, no. 9, pp. 1639–1645, 2009.

[52] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "Multiqc: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, 2016.

[53] L. Larcombe, R. Hendricusdottir, T. Attwood, F. Bacall, N. Beard, L. Bellis, W. Dunn, J. Hancock, A. Nenadic, C. Orengo, *et al.*, "Elixir-uk role in bioinformatics training at the national level and across elixir," *F1000Research*, vol. 6, 2017.

[54] J. J. Williams and T. K. Teal, "A vision for collaborative training infrastructure for bioinformatics," *Annals of the New York Academy of Sciences*, vol. 1387, no. 1, pp. 54–60, 2017.

[55] T. K. Attwood, S. Blackford, M. D. Brazas, A. Davies, and M. V. Schneider, "A global perspective on evolving bioinformatics and data science training needs," *Briefings in Bioinformatics*, vol. 20, pp. 398–404, Aug. 2017.

[56] "Community survey report - 2013 - embl-abr." https://www.embl-abr.org.au/news/braembl-community-survey-report-2013/.

[57] S. Roy, C. Coldren, A. Karunamurthy, N. S. Kip, E. W. Klee, S. E. Lincoln, A. Leon, M. Pullambhatla, R. L. Temple-Smolkin, K. V. Voelkerding, *et al.*, "Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the association for molecular pathology and the college of american pathologists," *The Journal of Molecular Diagnostics*, vol. 20, no. 1, pp. 4–27, 2018.

[58] M. Cavelaars, J. Rousseau, C. Parlayan, S. de Ridder, A. Verburg, R. Ross, G. R. Visser, A. Rotte, R. Azevedo, J.-W. Boiten, *et al.*, "Openclinica," in *Journal of clinical bioinformatics*, vol. 5, p. S2, Springer, 2015.

[59] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Reviews Genetics*, vol. 15, no. 6, pp. 409–421, 2014.

[60] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *cell*, vol. 100, no. 1, pp. 57–70, 2000.

[61] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell*, vol. 144, no. 5, pp. 646–674, 2011.

[62] Y. Lazebnik, "What are the hallmarks of cancer?," *Nature Reviews Cancer*, vol. 10, no. 4, pp. 232–233, 2010.

[63] S. D. Horne, S. A. Pollick, and H. H. Heng, "Evolutionary mechanism unifies the hallmarks of cancer," *International Journal of Cancer*, vol. 136, no. 9, pp. 2012–2021, 2015.

[64] Y. A. Fouad and C. Aanei, "Revisiting the hallmarks of cancer," *American journal of cancer research*, vol. 7, no. 5, p. 1016, 2017.

[65] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, p. 719, 2009.

[66] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C. R. Lane, E. P. Lim, N. Kalyanaraman, J. Nemesh, *et al.*, "Characterization of single-nucleotide polymorphisms in coding regions of human genes," *Nature genetics*, vol. 22, no. 3, p. 231, 1999.

[67] F. Zhang and J. R. Lupski, "Non-coding genetic variants in human disease," *Human molecular genetics*, vol. 24, no. R1, pp. R102–R110, 2015.

[68] S. Horn, A. Figl, P. S. Rachakonda, C. Fischer, A. Sucker, A. Gast, S. Kadel, I. Moll, E. Nagore, K. Hemminki, *et al.*, "Tert promoter mutations in familial and sporadic melanoma," *Science*, vol. 339, no. 6122, pp. 959–961, 2013.

[69] F. W. Huang, E. Hodis, M. J. Xu, G. V. Kryukov, L. Chin, and L. A. Garraway, "Highly recurrent tert promoter mutations in human melanoma," *Science*, vol. 339, no. 6122, pp. 957–959, 2013.

[70] A. de Klein, A. G. van Kessel, G. Grosveld, C. R. Bartram, A. Hagemeijer, D. Bootsma, N. K. Spurr, N. Heisterkamp, J. Groffen, and J. R. Stephenson, "A cellular oncogene is translocated to the philadelphia chromosome in chronic myelocytic leukaemia," *Nature*, vol. 300, no. 5894, p. 765, 1982.

[71] N. Heisterkamp, G. Jenster, J. Ten Hoeve, D. Zovich, P. K. Pattengale, and J. Groffen, "Acute leukaemia in bcr/abl transgenic mice," *Nature*, vol. 344, no. 6263, pp. 251–253, 1990.

[72] P. C. Nowell and D. A. Hungerford, "Chromosome studies on normal and leukemic human leukocytes," *Journal of the National Cancer Institute*, vol. 25, no. 1, pp. 85–109, 1960.

[73] P. C. Nowell and D. A. Hungerford, "Chromosome studies in human leukemia. ii. chronic granulocytic leukemia," *Journal of the National Cancer Institute*, vol. 27, no. 5, pp. 1013–1035, 1961.

[74] B. J. Druker, C. L. Sawyers, H. Kantarjian, D. J. Resta, S. F. Reese, J. M. Ford, R. Capdeville, and M. Talpaz, "Activity of a specific inhibitor of the bcr-abl tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the philadelphia chromosome," *New England Journal of Medicine*, vol. 344, no. 14, pp. 1038–1042, 2001.

[75] B. J. Druker, M. Talpaz, D. J. Resta, B. Peng, E. Buchdunger, J. M. Ford, N. B. Lydon, H. Kantarjian, R. Capdeville, S. Ohno-Jones, *et al.*, "Efficacy and safety of a specific inhibitor of the bcr-abl tyrosine kinase in chronic myeloid leukemia," *N Engl j Med*, vol. 2001, no. 344, pp. 1031–1037, 2001.

[76] K. Yi and Y. S. Ju, "Patterns and mechanisms of structural variations in human cancer," *Experimental & molecular medicine*, vol. 50, no. 8, pp. 1–11, 2018.

[77] M. N. H. Luijten, J. X. T. Lee, and K. C. Crasta, "Mutational game changer: Chromothripsis and its emerging relevance to cancer," *Mutation Research/Reviews in Mutation Research*, vol. 777, pp. 29–51, jul 2018.

[78] J. Yang, J. Liu, L. Ouyang, Y. Chen, B. Liu, and H. Cai, "CTLPScanner: a web server for chromothripsis-like pattern detection," *Nucleic Acids Research*, vol. 44, pp. W252–W258, may 2016.

[79] S. K. Govind, A. Zia, P. H. Hennings-Yeomans, J. D. Watson, M. Fraser, C. Anghel, A. W. Wyatt, T. van der Kwast, C. C. Collins, J. D. McPherson, R. G. Bristow, and P. C. Boutros, "ShatterProof: operational detection and quantification of chromothripsis," *BMC Bioinformatics*, vol. 15, no. 1, p. 78, 2014.

[80] C. Swanton, "Intratumor heterogeneity: evolution through space and time," *Cancer research*, vol. 72, no. 19, pp. 4875–4882, 2012.

[81] F. Pacchierotti and M. Spanò, "Environmental impact on dna methylation in the germline: state of the art and gaps of knowledge," *BioMed research international*, vol. 2015, 2015.

[82] J. O'Rawe, T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W. E. Johnson, Z. Wei, K. Wang, and G. J. Lyon, "Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing," *Genome Medicine*, vol. 5, no. 3, p. 28, 2013.

[83] S. Hwang, E. Kim, I. Lee, and E. M. Marcotte, "Systematic comparison of variant calling pipelines using gold standard personal exome variants," *Scientific reports*, vol. 5, p. 17875, 2015.

[84] J. M. Zook, B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide, and M. Salit, "Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls," *Nature biotechnology*, vol. 32, no. 3, pp. 246–251, 2014.

[85] H. Xu, J. DiCarlo, R. V. Satya, Q. Peng, and Y. Wang, "Comparison of somatic mutation calling methods in amplicon and whole exome sequence data," *BMC genomics*, vol. 15, no. 1, p. 244, 2014.

[86] S. Y. Kim and T. P. Speed, "Comparing somatic mutation-callers: beyond venn diagrams," *BMC bioinformatics*, vol. 14, no. 1, p. 189, 2013.

[87] N. D. Roberts, R. D. Kortschak, W. T. Parker, A. W. Schreiber, S. Branford, H. S. Scott, G. Glonek, and D. L. Adelson, "A comparative analysis of algorithms for somatic snv detection in cancer," *Bioinformatics*, vol. 29, no. 18, pp. 2223–2230, 2013.

[88] T. S. Alioto, I. Buchhalter, S. Derdak, B. Hutter, M. D. Eldridge, E. Hovig, L. E. Heisler, T. A. Beck, J. T. Simpson, L. Tonon, *et al.*, "A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing," *Nature communications*, vol. 6, p. 10001, 2015.

[89] A. B. Krøigård, M. Thomassen, A.-V. Lænkholm, T. A. Kruse, and M. J. Larsen, "Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data," *PloS one*, vol. 11, no. 3, p. e0151664, 2016.

[90] M. Callari, S.-J. Sammut, L. D. Mattos-Arruda, A. Bruna, O. M. Rueda, S.-F. Chin, and C. Caldas, "Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers," *Genome Medicine*, vol. 9, apr 2017.

[91] V. Vijayan, S.-M. Yiu, and L. Zhang, "Improving somatic variant identification through integration of genome and exome data," *BMC Genomics*, vol. 18, oct 2017.

[92] F. J. Sedlazeck, A. Dhroso, D. L. Bodian, J. Paschall, F. Hermes, and J. M. Zook, "Tools for annotation and comparison of structural variation," *F1000Research*, vol. 6, p. 1795, oct 2017.

[93] S. W. Scherer, C. Lee, E. Birney, D. M. Altshuler, E. E. Eichler, N. P. Carter, M. E. Hurles, and L. Feuk, "Challenges and standards in integrating surveys of structural variation," *Nature Genetics*, vol. 39, pp. S7–S15, jul 2007.

[94] F. J. Sedlazeck, P. Rescheneder, M. Smolka, H. Fang, M. Nattestad, A. von Haeseler, and M. C. Schatz, "Accurate detection of complex structural variations using single-molecule sequencing," *Nature Methods*, vol. 15, pp. 461–468, apr 2018.

[95] J. D. Merker, A. M. Wenger, T. Sneddon, M. Grove, Z. Zappala, L. Fresard, D. Waggott, S. Utiramerur, Y. Hou, K. S. Smith, S. B. Montgomery, M. Wheeler, J. G. Buchan, C. C. Lambert, K. S. Eng, L. Hickey, J. Korlach, J. Ford, and E. A. Ashley, "Long-read genome sequencing identifies causal structural variation in a mendelian disease," *Genetics in Medicine*, vol. 20, pp. 159–163, jun 2017.

[96] X. Fan, M. Chaisson, L. Nakhleh, and K. Chen, "Hysa: a hybrid structural variant assembly approach using next-generation and single-molecule sequencing technologies," *Genome research*, vol. 27, no. 5, pp. 793–800, 2017.

[97] M. H. Weissensteiner, A. W. Pang, I. Bunikis, I. Höijer, O. Vinnere-Petterson, A. Suh, and J. B. Wolf, "Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications," *Genome Research*, vol. 27, pp. 697–708, mar 2017.

[98] J. R. Miller, P. Zhou, J. Mudge, J. Gurtowski, H. Lee, T. Ramaraj, B. P. Walenz, J. Liu, R. M. Stupar, R. Denny, L. Song, N. Singh, L. G. Maron, S. R. McCouch, W. R. McCombie, M. C. Schatz, P. Tiffin, N. D. Young, and K. A. T. Silverstein, "Hybrid assembly with long and short reads improves discovery of gene family expansions," *BMC Genomics*, vol. 18, jul 2017.

[99] A. Ritz, A. Bashir, S. Sindi, D. Hsu, I. Hajirasouliha, and B. J. Raphael, "Characterization of structural variants with single molecule and hybrid sequencing approaches," *Bioinformatics*, vol. 30, pp. 3458–3466, oct 2014.

[100] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "The human microbiome project," *Nature*, vol. 449, no. 7164, p. 804, 2007.

[101] S. D. Ehrlich, M. Consortium, *et al.*, "Metahit: The european union project on metagenomics of the human intestinal tract," in *Metagenomics of the human body*, pp. 307–316, Springer, 2011.

[102] R. Sender, S. Fuchs, and R. Milo, "Are we really vastly outnumbered? revisiting the ratio of bacterial to host cells in humans," *Cell*, vol. 164, pp. 337–340, jan 2016.

[103] E. A. Grice and J. A. Segre, "The human microbiome: Our second genome," *Annual Review of Genomics and Human Genetics*, vol. 13, pp. 151–170, sep 2012.

[104] I. Cho and M. J. Blaser, "The human microbiome: at the interface of health and disease," *Nature Reviews Genetics*, vol. 13, pp. 260–270, mar 2012.

[105] Z. Gao, C.-h. Tseng, B. E. Strober, Z. Pei, and M. J. Blaser, "Substantial alterations of the cutaneous bacterial biota in psoriatic lesions," *PloS one*, vol. 3, no. 7, p. e2719, 2008.

[106] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon, "An obesity-associated gut microbiome with increased capacity for energy harvest," *nature*, vol. 444, no. 7122, p. 1027, 2006.

[107] R. E. Ley, F. Bäckhed, P. Turnbaugh, C. A. Lozupone, R. D. Knight, and J. I. Gordon, "Obesity alters gut microbial ecology," *Proceedings of the National Academy of Sciences*, vol. 102, no. 31, pp. 11070–11075, 2005.

[108] M. Castellarin, R. L. Warren, J. D. Freeman, L. Dreolini, M. Krzywinski, J. Strauss, R. Barnes, P. Watson, E. Allen-Vercoe, R. A. Moore, *et al.*, "Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma," *Genome research*, vol. 22, no. 2, pp. 299–306, 2012.

[109] A. D. Kostic, D. Gevers, C. S. Pedamallu, M. Michaud, F. Duke, A. M. Earl, A. I. Ojesina, J. Jung, A. J. Bass, J. Tabernero, *et al.*, "Genomic analysis identifies association of fusobacterium with colorectal carcinoma," *Genome research*, vol. 22, no. 2, pp. 292–298, 2012.

[110] R. Ranjan, A. Rani, A. Metwally, H. S. McGee, and D. L. Perkins, "Analysis of the microbiome: Advantages of whole genome shotgun versus 16s amplicon sequencing," *Biochemical and biophysical research communications*, vol. 469, no. 4, pp. 967–977, 2016.

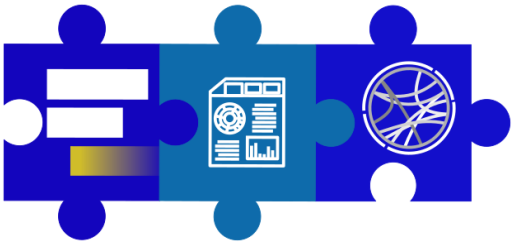
[111] J.-C. Lagier, P. Hugon, S. Khelaifia, P.-E. Fournier, B. La Scola, and D. Raoult, "The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota," *Clinical microbiology reviews*, vol. 28, no. 1, pp. 237–264, 2015.

[112] S. Yang and R. E. Rothman, "Pcr-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings," *The Lancet infectious diseases*, vol. 4, no. 6, pp. 337–348, 2004.

[113] A. M. Caliendo, D. N. Gilbert, C. C. Ginocchio, K. E. Hanson, L. May, T. C. Quinn, F. C. Tenover, D. Alland, A. J. Blaschke, R. A. Bonomo, *et al.*, "Better tests, better care: improved diagnostics for infectious diseases," *Clinical Infectious Diseases*, vol. 57, no. suppl_3, pp. S139–S170, 2013.

[114] C. Song and H. Chen, "Overview of research on fusion genes in prostate cancer," *TRANSLATIONAL CANCER RESEARCH*, vol. 9, no. 3, pp. 1998–+, 2020.

[115] K. M. Koo, P. N. Mainwaring, S. A. Tomlins, and M. Trau, "Merging new-age biomarkers and nanodiagnostics for precision prostate cancer management," *Nature Reviews Urology*, vol. 16, no. 5, pp. 302–317, 2019.



"I was taught that the way of progress was neither swift nor easy."

Marie Curie

1

Accessible Bioinformatics

The vast majority of bioinformatics tools are UNIX-based command line applications that are not easy to use for non-bioinformaticians. This chapter describes three resources developed to increase the accessibility of bioinformatics tools and pipelines for non-experts.

The **Galaxy platform** aims to improve the user-friendliness of bioinformatics tools by providing a web-based graphical user interface to command line UNIX tools. This enables data scientists to

run complex bioinformatics analyses needing nothing more than a web browser.

Analysis pipelines often generate a large number of outputs, and with the explosion of data in next generation sequencing field, many of these resulting outputs contain millions of lines and can no longer easily be reviewed manually.

Circos is a tool that enables visualisation of high-dimensional data in a circular plot. Circos is extremely flexible but has a high degree of complexity and a steep learning curve. We developed **Galactic Circos**, a Galaxy version of the Circos tool, to increase its usability for non-bioinformaticians, and to improve interoperability with upstream tools.

Visualisation and summarisation of results has become an integral part of any analysis pipeline. The **iReport** tool provides generic reporting within the Galaxy framework for easy summarization of analysis results. This result can help researchers and clinician quickly interpret their data.

This chapter contains the following sub-chapters:

1.0 **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update**

The Galaxy platform is an enormous effort supported by a very big team of developers, scientists and community members. My contributions to this work consisted primarily of the following activities. 1) Developing tools (over 150 tool including mothur, cgatools, iReport, Circos, and many more) and workflows. Furthermore, as member of the IUC (Intergalactic Utilities Commission) I was involved in developing best-practice guidelines for tool development and was actively involved in community building efforts around the tools. 2) Training, as described in more detail in the next chapter, I co-founded the Galaxy Training Materials project, an effort to centralized and FAIR-ify training materials that use the Galaxy platform. 3) I created several *interactive environments* in Galaxy; unlike most tools, these special container-based tools allow for interactive exploration of data. I created interactive environments for Phinch (visualization of microbiome data), Ethercalc (a spreadsheet application), and contributed to the Rstudio interactive environment. I also made some (minor) contributions to the infrastructure supporting interactive environments in Galaxy. 4) I created a connector between the Owncloud data sharing platform (`https://owncloud.com/`) and Galaxy, adding a "Send to Galaxy" button on datasets in Owncloud, and vice versa. 5) I've co-created Galaxy *flavours*, most notably for metagenomics. These docker-based topical Galaxy instances provide ready-to-deploy

Galaxy containers optimized for a specific research domain. 6) As an administrator of a (small) public Galaxy server, I've contributed bug fixes to the Galaxy base code over the years as well. Furthermore, I often provide feedback about the requirements of researchers in order to guide the roadmap of Galaxy developments and ensure these are in line with the needs of the community.

1.1 **Galactic Circos: User-friendly creation of Circos Plots within the Galaxy platform.** Circos is a very popular tool for visualisation of genome-scale data, especially in publications. While Circos is incredibly powerful and flexible, it is also complex to the point of requiring bioinformatics expertise to use. Helena Rasche and I created a Galaxy port of this tool, so that Circos plots can be configured by users via the browser. Due to the complexity of Circos, this became probably the single-most complex Galaxy tool wrapper as well. While not exposing the full power of Circos, it strikes a good balance between versatility and complexity, between flexibility and user-friendliness. Because of the complexity of this tool, we also created detailed training materials as part of the publication in order to increase the accessibility of the tool.

1.2 **iReport: A generalised Galaxy solution for integrated experimental reporting.** I developed this Galaxy tool as a way to offer generic, configurable analysis results reporting within Galaxy. These reports are crucial for clinical applications where the most important analysis results need to be presented in a succinct and sharable report. Note that this paper refers to the CTMM-TraIT Tool Shed, which has since been taken offline. However, The iReport tool is now available in the main Galaxy Tool Shed. Furthermore, in the years since I wrote iReport, the Galaxy team has recognized the importance of this tool, and has worked towards incorporating this functionality into the Galaxy code base itself. Termed *Galaxy workflow reports*, an iReport-like results report can now be configured into the Galaxy workflow definition itself. While iReport is still available from the Tool Shed and can still be used, I have halted any further iReport development in favour of contributing to these Galaxy workflow reports and optimizing their utility in a research and clinical setting.

The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update

Enis Afgan[7,*], Dannon Baker[7], Bérénice Batut[1,*], Marius van den Beek[3,*], Dave Bouvier[6,*], Martin Čech[6,*], John Chilton[6,*], Dave Clements[7,*], Nate Coraor[6,*], Björn Grüning[1,*], Aysam Guerler[7,*], Jennifer Hillman-Jackson[6,*], Saskia Hiltemann[1,*], Vahid Jalili[7,*], Helena Rasche[1,*], Nicola Soranzo[14], Jeremy Goecks[7,*], James Taylor[7,*], Anton Nekrutenko[6,*], Daniel Blankenberg[9,*]

* The authors wish it to be known that, in their opinion, all authors should be regarded as Joint First Authors.

Published in: *Nucleic Acids Research*, Volume 46, Issue W1, 2 July 2018, Pages W537–W544, DOI: https://doi.org/10.1093/nar/gky379

1. Department of Biology, Johns Hopkins University, Baltimore MD USA.
2. Department of Comuter Science, Albert-Ludwigs-University, Freiburg, Germany.
3. Institut Curie, PSL Research University, Paris, France.
4. Department of Biochemistry and Molecular Biology, Penn State University, University Park, PA, USA.
5. Center for Biological Systems Analysis (ZBSA), University of Freiburg, Freiburg, Germany.
6. Department of Pathology, Erasmus Medical Centre, Rotterdam, The Netherlands.
7. Department of Biomedical Engineering, Oregon Health and Science University, OR, USA.
8. Earlham Institute, Norwich Research Park, Norwich, United Kingdom.
9. Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA.

Abstract

Galaxy (homepage: https://galaxyproject.org, main public server: https://usegalaxy.org) is a web-based scientific analysis platform used by tens of thousands of scientists across the world

to analyze large biomedical datasets such as those found in genomics, proteomics, metabolomics, and imaging. Started in 2005, Galaxy continues to focus on three key challenges of data-driven biomedical science: making analyses accessible to all researchers, ensuring analyses are completely reproducible, and making it simple to communicate analyses so that they can be reused and extended. During the last two years, the Galaxy team and the open-source community around Galaxy have made substantial improvements to Galaxy's core framework, user interface, tools, and training materials. Framework and user interface improvements now enable Galaxy to be used for analyzing tens of thousands of datasets, and more than 5,500 tools are now available from the Galaxy ToolShed. The Galaxy community has led an effort to create numerous high-quality tutorials focused on common types of genomic analyses. The Galaxy developer and user communities continue to grow and be integral to Galaxy's development. The number of Galaxy public servers, developers contributing to the Galaxy framework and its tools, and users of the main Galaxy server have all increased substantially.

Introduction

Advances in biomedicine and biology increasingly rely on analysis of large datasets. Started in 2005, the Galaxy Project (https://galaxyproject.org) [0, 1, 2] maintains a focus on enabling data-driven biomedical science by pursuing three goals: (a) accessible data analysis serving all scientists regardless of their informatics expertise and tool developers seeking a wider audience and broad integration of their tools; (b) reproducible analyses regardless of the particular platform; and (c) transparent communication of analyses, which in turn enables reuse and extension of analyses across communities of practice. The Galaxy Project consists of four complementary components:

1. The main public Galaxy server (https://usegalaxy.org) — this server is the subject of this article and has been online since 2007. It features a rich toolset for large-scale genomics analyses, terabytes of public data for use, and hundreds of shared analysis histories, workflows, and interactive publication supplements. This server has more than 124,000 registered users whom run 245,000 analysis jobs each month.

2. The Galaxy framework and software ecosystem (https://github.com/galaxyproject) — an open-source software package that anyone can use to run a Galaxy server on any Unix-based operating system. The Galaxy ecosystem includes a software development kit (SDK) for Galaxy tool development, API language bindings for multiple programming languages, software for scripting Galaxy interactions, and tools for automating setup and deployment of Galaxy and its plugins such as tools and visualizations.

3. The Galaxy ToolShed (https://toolshed.g2.bx.psu.edu/) — a community-driven resource for the dissemination of Galaxy tools, workflows, and visualizations. This server functions as an "AppStore" for Galaxy servers where developers and Galaxy admins can host, share, and install Galaxy tools, workflows, and visualizations.

4. The Galaxy Community (https://galaxyproject.org/community/) — distinct and complementary subcommunities make key contributions to all aspects of the Project. These subcommunities address the needs and desires of every category of stakeholder including users, administrators, developers, resource providers, and educators.

Galaxy has served hundreds of thousands of users, been used in more than 5,700 scientific publications, and provided 500+ developers with a framework provisioning accessible, transparent, and reproducible data analysis (https://galaxyproject.org/galaxy-project/statistics/). Many instances of the framework have been installed, including Galaxy Main (https://usegalaxy.org) and over 99 publicly accessible servers (https://galaxyproject.org/public-galaxy-servers/), serving biomedical and other domain-specific research. Significant growth has occurred across all sectors of the Galaxy Project within the past two years (Fig. 1.0).

New Features

Scalability

Scalability is amongst the most significant challenges that Galaxy faces as the size and number of biomedical and especially genomics datasets continues to grow. For instance, single-cell RNA-seq experiments routinely generate hundreds or thousands of primary datasets. As a web-based application, Galaxy must scale both in its web-based interface and on its backend server and do so in a multiuser environment.

User interface scalability enables scientists to use the Galaxy web interface to analyze many datasets, apply (collective) operations on them, and design pipelines to analyze them. Galaxy implements a variety of features to facilitate analyzing large numbers of datasets, including workflows and collections. Our recent optimizations of the user interface (UI) yielded a significant improvement to frontend scalability. We benchmarked the optimizations by replicating an experiment conducted on single Hematopoietic stem cells and multipotent progenitors [3] to quantify the expression of 64,000 transcripts, which generates 11,872 history items. Galaxy ran this

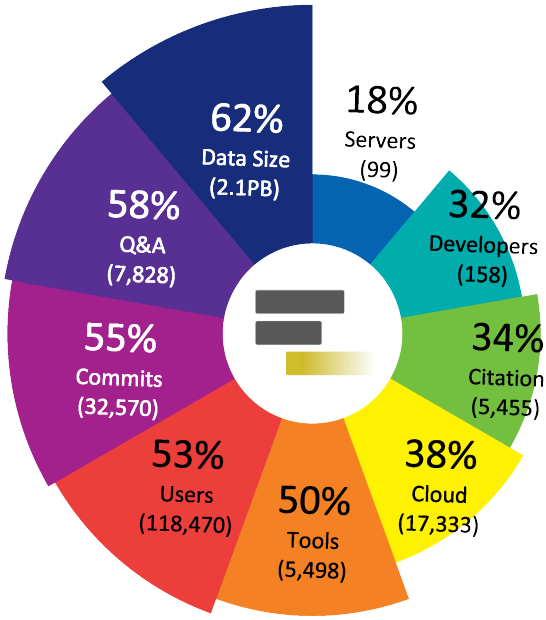


Figure 1.0: Circular barplot illustrating recent growth of the Galaxy Project across several independent facets. In the past two years, usage of the main public Galaxy server has increased 60%, the number of tools and supported versions has increased 53%, and the amount of data analyzed on the main server has increased 72%. A growing number of public instances (18% increase) and cloud-based Galaxies (38% increase) provide researchers with a wider range of options for scalability and application domains. Additionally, more developers (45% increase with 63% more commits to the codebase) contributed to the Galaxy framework and software ecosystem. Question and answer activity on the Galaxy Biostars forum increased 68%.

proof of concept experiment seamlessly using existing standard tools, whereas earlier versions of Galaxy would not have been able to support this analysis.

Server scalability refers to the Galaxy's ability to execute many data analysis/manipulation tasks for many users. This is achieved by advantageously utilizing a range of available computing resources. The Galaxy framework runs on various platforms, from a standard laptop to institutional clusters and cloud-based platforms. Galaxy is highly versatile in its ability to deploy jobs (atomic units of work), as it can leverage a multitude of workload managers including Slurm [4], HTCondor [5], Apache Mesos [6], and Kubernetes (`https://kubernetes.io`), among others, in addition to a built-in lightweight job running system. Recent enhancements to Galaxy's job management include dynamic job destination assignment (which facilitate automatic job parameter-specific resource selection), delay in job queuing (e.g., for workflows), automatic job re-submission (e.g., on job failure due to a temporary cluster error), and means of implementing fair-share prioritization schemes. These features are being used on Galaxy Main (Fig. 1.1) to leverage

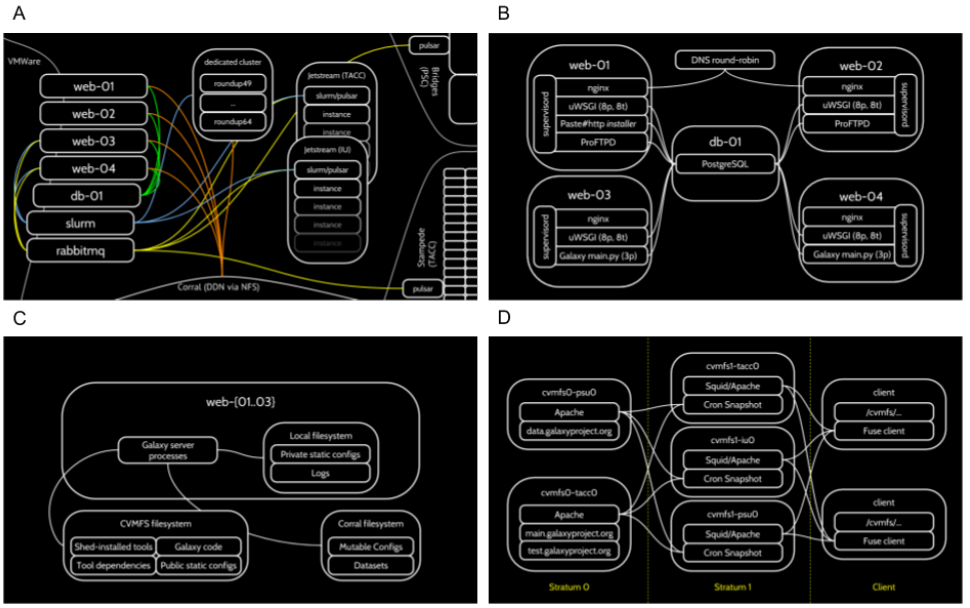


Figure 1.1: Schematic of servers and services in use at Galaxy Main. (A) A global overview of Galaxy Main resources. (B) Multiple frontend servers serve Galaxy content to users by utilizing round-robin load balancing. (C) Layout of data schemes used by Galaxy Main is optimized for application speed, concurrent access, and versioned content. (D) CVMFS infrastructure hosted by the Galaxy Project that is used at Main and available for access to any other Galaxy instance.

cloud computing resources for better job throughput. Specifically, Galaxy Main is now configured to take advantage of the XSEDE infrastructure [7] that includes Bridges and Stampede resources as well as the Jetstream cloud [8]. The benefits of using these resources include the ability to run larger jobs, as shown in Figure 1.2. Additionally, use of these resources has enabled new types of analysis to be enabled on Main. Notably, this includes Galaxy Interactive Environments through to the ability to use containerization technologies and provide sufficient isolation of individual jobs from other processes running on the same underlying compute infrastructure.

A complete Galaxy server with a full repertoire of tools and reference data can be run on major cloud platforms. These servers are launched independently by users, and come pre-configured with hundreds of tools and reasonable default settings typical of a production server. Notably, launched instances do not have usage quotas and can be customized to install any desired tool. We have designed a cloud-agnostic approach for leveraging these resources by developing the abstraction

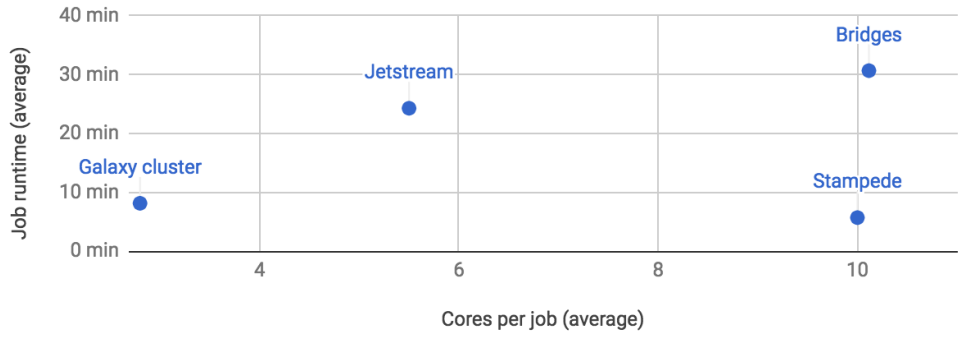


Figure 1.2: Enabling automated selection and use of specialized national cyberinfrastructure compute resources from Galaxy Main enhances user-experience. It is now possible to run jobs that are up to an order of magnitude larger than before by using Bridges and Stampede. New types of jobs, such as interactive environments (see Advances in tools section), that require execution isolation due to security concerns are enabled by utilizing virtualization facilitated by the Jetstream cloud. Consequently, it is possible to concurrently run more jobs due to the increase in processing capacity.

library CloudBridge [9] and a new CloudLaunch application. These two solutions make it possible to launch Galaxy instances across a variety of cloud providers while reducing the requirement to build and maintain cloud-specific resources (e.g., machine images, file systems). There are now ten different flavors of Galaxy available for launching on major clouds including Amazon Web Services, Jetstream, and Microsoft Azure (`https://launch.usegalaxy.org`).

Advances in tools

The Galaxy ToolShed [10] assumes the role of an AppStore for Galaxy instances by hosting thousands of tools. The ToolShed improves tool availability, deployment, and portability across Galaxy servers and computing environments.

Updated tool suite Over the last two years, we have expanded both the quantity and quality of the tools available on the Galaxy Toolshed. As of April 2018, the ToolShed hosts 5,628 tools, which shows 53% growth since 2016, and approximately 2,000 repositories had at least one new update. Examples of new tools include: GEMINI for exploring genetic variation [11]; mothur for analyzing rRNA gene sequences [12]; QIIME for quantitative microbiome analysis from raw DNA sequencing data [13]; deepTools for explorative analysis of deeply sequence data [14, 15]; HiCexplorer [16] for analysis and visualization of Hi-C data; ChemicalToolBox for comprehensive access to cheminformatics libraries and drug discovery tools [17]; minimap2 (`https://arxiv.org/abs/1708.01492`) and poretools for long read sequencing analysis [18]; MultiQC [19] to aggregate

multiple results into a single report; a new RNA-seq analysis tool suite with modern analysis tools such as Kallisto [20], Salmon [21], Deseq2 [22], and STAR-Fusion [23]; and GenomeSpace [24], a cloud-based interoperability tool.

Tool environment and interface. The portability and backward-compatibility of the Galaxy tools environment is improved significantly. Accordingly, a tool configuration now includes a tool profile version, which is used to ensure compatibility between a version of a tool and its targeted Galaxy version. In addition, tool profile versions allow for the evolution of new and better tool defaults and behaviors while maintaining backwards compatibility. We also improved the ToolShed API and its interface to facilitate installing tools missing from an imported workflow. We improved the installation process so that restarting Galaxy is not required to use a newly installed tool.

Interactive analysis and visualization

Galaxy's UI makes it possible for anyone to run complex analyses. However, a complete analysis of genomic data often requires custom scripts or visualizations, especially at the beginning (data preparation) or end (data summarization) of analyses. To meet these customized needs, we recently introduced Galaxy Interactive Environments [25], an integration of Galaxy with Jupyter (RStudio is in development)—a commonly used interactive scripting platform. With Interactive Environments, Galaxy users benefit from existing computational infrastructure via both graphical UI and ad hoc scripting, or any combination of these.

Galaxy's visualization framework [26] makes it possible to integrate a wide variety of Web-based and server-side visualizations. Through this framework, many new visualizations have been added to Galaxy, including Cytoscape [27], and the WebGL enabled 3D Protein viewer NGL [28], molecular interaction networks and macromolecular structures visualizations, and the 100+ visualizations available through BioJS [29], a rich set of community-driven JavaScript components for agile and interactive visualization of biological data.

User interface and experience enhancements

There are two common modes of data analysis: exploratory and pipeline execution. Galaxy enables simultaneous access to both of these. Users are able to interactively analyze their data by making use of individual tools in a trial-and-error manner. They are then able to automatically generate reusable and generalizable workflows from an ad hoc analysis. An interactive workflow

editor is also available to modify or generate workflows from scratch. At any point in time, a user can seamlessly switch modes between interactively analyzing a datasets and executing a workflow on these datasets. There is no analysis lock-in, and users can exercise full control, or make use of pre-existing pipelines. Importantly, these analysis artefacts, such as datasets, analysis histories, workflows, and visualizations can all be shared and copied by collaborators at the discretion of the analyst.

Client-side infrastructure The client-side of Galaxy, which is the user-interface most people associate Galaxy with, has seen significant changes under the hood. The usage of server-side mako templates, for example to create forms, has been further reduced and replaced by client-side only code that communicates via the RESTful Galaxy API with the backend. This minimizes the number of full-page refreshes and improves response time by enabling partial page updates. The interface has been further enhanced to allow for drag-and-drop of files and datasets, presents a fuzzy search on dataset and tool metadata, and implements a modal scratchbook for visualizations and comparison of multiple datasets.

Furthermore, the community has selected the Vue.js framework (`https://vuejs.org/`) as the base for future improvements allowing all UI elements to converge into a more reactive and future-proof interface. With the integration of Vue.js, the entire client-side build system was updated to utilize the latest web-technologies, to make routing and loading times faster, and to encourage rapid future interface improvements. While mostly transparent to users, these changes are the fundamental groundwork of a more more flexible UI framework that will enable visual enhancements and an improved user experience for years to come.

Tags. Although tags have been supported in Galaxy for several years, they have only recently become advantageous for large many-sample analyses. We have enhanced tags to allow propagation through dataset analysis steps. This facilitates tracking individual datasets through the entire analysis life-cycle and becomes part of the provenance system and ease-of-use of Galaxy. To enable automatic tag propagation, a hash-sign (#) is placed at the beginning of the tag, which is colloquially referred to as a named-tag. While standard Galaxy output dataset naming is suitable for many interactive analyses, the connection between inputs and outputs through large workflows becomes increasingly less obvious; by utilizing named-tags, users can label datasets with an identifier that is maintained throughout the analysis.

Webhooks. Inspired by user feedback and the need to quickly modify and adapt Galaxy's interface, we integrated a pluggable system to extend Galaxy's frontend. Webhooks provide an entry-point into the Galaxy UI, in which it is possible to add buttons, menu entries, or entire iframes. At these entry-points a developer can dynamically add client-side code (JavaScript, HTML, CSS) and interact with the rest of the Galaxy user-interface. By integrating Webhooks with the Galaxy API, it is also possible to trigger server-side functions from within a Webhook. Webhooks can be thoroughly customized and are enabled at the discretion of the Galaxy administrator.

Interactive tours. We have developed self-paced, interactive tours that users can step through to learn about Galaxy. These tours guide users step by step through using the interface including tools, workflows, and other features available in Galaxy. To simplify tour creation, a Tour Builder (`https://github.com/TailorDev/galaxy-tourbuilder`) has been created for recording, replaying, updating, and exporting tours.

Improved workflows. Galaxy workflows have been extended in several ways. Switching between tool versions and upgrading workflows with new tool versions is now supported. A workflow can now be embedded in another, making it easier to create and edit workflows that have many common steps repeated. Many of these features have existed in in standalone workflow systems, such as Taverna [30], for sometime, but have been widely requested by Galaxy users. Workflows are now scheduled by a Galaxy server more efficiently and in the background, making it possible to execute larger workflows, generating tens of thousands of jobs, while providing instant feedback and a snappier user-experience. We have also enhanced Galaxy with initial support for running workflows defined in the Common Workflow Language [31] format.

Dataset collections Galaxy Dataset Collections combine datasets to enable simultaneous analysis. They organize sets of datasets as potentially nested lists of objects allowing easier data handling and batch execution of tools. In addition to the related frontend improvements, and support of nesting collections together, we recently introduced specialized tools to be executed on collections (e.g., Collapse, which combines a list of datasets into a single dataset, Flatten which takes nested collections and produces a flat list of datasets, and Merge which takes two lists and creates a single unified list), and enabled uploading and downloading dataset collections to and from both user's local disk and Galaxy data libraries.

Infrastructure enhancements

In order to make Galaxy more robust in a production environment, we adopted technologies to enhance Galaxy's portability, security, reliability, and scalability. Galaxy now utilizes uWSGI (http://projects.unbit.it/uwsgi) as its default web application server. This adoption has several advantages, namely the ability to negate Python's limitations regarding concurrent tasks execution, built-in load balancing, scalability, improved fault tolerance, and the possibility of restarting Galaxy uninterruptedly.

Many tools available via Galaxy rely on the availability of reference and index data. To promote ease of use and efficient storage and compute resources, Galaxy is able to share a precomputed set of local reference data for tools to use. Previously, making this data available to the tools was a time intensive process where a Galaxy administrator had to install and properly configure the server, either manually or by using Data Managers [32]. However, this resulted in much redundant effort required for each Galaxy server being configured. To streamline this process, we have made all the reference data we prepared for Galaxy Main available via a CernVM File System [33], a scalable and content-addressable file system. This repository currently hosts 5TB of pre-build reference data, which are versioned and shared publicly with read-only access. With minimal configuration, any instance of Galaxy, including Galaxy-Docker images, can attach to this file system and gain access to the same reference data available on Galaxy Main. To improve accessibility and fault-tolerance, this data source is replicated on servers located in Europe and Australia.

Galaxy is powered by various open-source projects which are installed automatically, and used when needed. Galaxy is using the Conda package manager (https://conda.io) as its default tool dependency resolver, and offers support for virtualization and containerization technologies (e.g., Docker (https://www.docker.com) and Singularity [34]) to ensure a higher level of portability, if needed. By leveraging the Bioconda (https://doi.org/10.1101/207092) and the BioContainer [35] projects, Galaxy is able to provision and use reproducible tool execution environments.

Galaxy is a generic data analysis framework, which can be configured for various application scenarios using a wide range of configuration parameters. To facilitate configuring these parameters with optimal values for a number of predefined application scenarios, the Galaxy project leverages Ansible (https://www.ansible.com), software for automated configuration and management of other software packages. We have developed and shared Ansible configurations for Galaxy Main, the main public Galaxy server, (https://github.com/galaxyproject/usegalaxy-playbook) and also a configurable generic playbook for setting up production instances on cloud resources,

virtual machines, and bare metal (https://github.com/ARTbio/GalaxyKickStart). This playbook can be used as a reference for configuring a Galaxy instance for a production environment.

The Galaxy-Docker project (https://github.com/bgruening/docker-galaxy-stable), delivers a production ready Galaxy instance in minutes and can be used as the basis for personalised, self-contained, portable instances of Galaxy, known as Galaxy flavors. Preconfigured by the Galaxy community a plenitude of flavors already exist covering application scenarios, from BLAST+ [36, 37], metagenomics (https://doi.org/10.1101/183970), ChIP-exo analysis, or RNA research [38]. In addition to the facilitated and out-of-box functionality, these images provision isolated environments well-suited for experimenting with tools and Galaxy configurations, and are ideal for training courses, as demonstrated by the Galaxy Training Network.

Server monitoring and issue management is crucial in production Galaxy instances. Galaxy has integrated a plugin module to submit user bug-reports to configurable endpoints such as mailing lists or GitHub issues. With this, Galaxy can be configured to send error reports to a local ticket system. The recent integration of Sentry (https://sentry.io/) for automated error tracking and reporting makes it easier for administrators to track both client- and server-side errors without requiring manual user bug reports.

Community

Galaxy serves several distinct communities: researchers, tool developers, resource providers, trainers, and trainees. To centralize resources for all communities, we have developed the The Galaxy Community Hub (https://galaxyproject.org) for all things Galaxy. The Hub uses a modified wiki approach, with content written in Markdown, a simple formatting language, and then built into a static website. Anyone can update the Markdown documents using GitHub pull requests, a standard approach for collaborating on code and documentation on GitHub projects. Submitted pull requests are reviewed and merged, and the Hub site is automatically regenerated and updated, resulting in high-quality reviewed content that can be updated by any member of the Galaxy community. The Hub includes a full list of public Galaxy servers (https://galaxyproject.org/public-galaxy-servers), a large set of tutorials for learning to use Galaxy and perform genomic analyses, extensive documentation on deploying and administering a Galaxy server in the Cloud or on local hardware, and upcoming events. We also maintain an annotated listing of the more than 5,000 publications referencing Galaxy via the free and open-source Zotero service (https://www.zotero.org/groups/1732893/galaxy).

The Main Galaxy server has over 124,000 registered users and approximately 2,000 new users register each month. On average, 20,000 unique users execute over 245,000 analysis jobs by accessing 750 different tools every month. With such an active user-base, questions on platform and tool usage, as well as general research questions [39], are common. To efficiently assist users in performing research, we provide a Biostars [40] Question and Answers forum (https://biostar.usegalaxy.org/) that leverages the knowledge and strength of community members to provide support. This forum is monitored and moderated by core team members, but the Galaxy user community provides many answers. Help is also available through live chat with the team and community members via Gitter and IRC chat services, which are used most often by developers and administrators. In addition to the online help and documentation, the Galaxy Training Network has developed comprehensive tutorials and workflows for performing common data analysis tasks, providing topic-specific introduction slides, hands-on material, sample data, and even playable Galaxy tours (https://doi.org/10.1101/225680).

Many in-person events that highlight and build the Galaxy community occur each year (https://galaxyproject.org/events/). These include free or low-cost hands-on workshops and training sessions that have been hosted by the community on six continents. The Galaxy Community Conference (GCC) is an annual conference that was first held in 2010. GCC alternates between Europe and the United States, includes two full days of training, two days of coding and data analysis hackathons, and two days of oral and poster presentations. Galaxy conferences have had over two hundred attendees each year since 2012, and over eleven hundred different researchers have attended since 2010. Our 2018 conference will be hosted jointly with the Bioinformatics Open Source Conference (BOSC) in an effort to promote and centralize discussion of open-source software for bioinformatics.

Another core area of community focus is tool development and availability. The Intergalactic Utilities Commision (IUC; https://galaxyproject.org/iuc/) is a community-based organization that defines best-practices for tool development that help ensure the availability of high-quality tools in the ToolShed. It is a self-organizing and self-regulating group that has grown by six new members in the last two years and is primarily composed of individuals outside of the core Galaxy development team. The IUC is only one of many tool contributors, with the ToolShed allowing any member of the community to share tools that they have added to Galaxy. To assist community members with tool development and distribution, a command line tool named Planemo (https://github.com/galaxyproject/planemo) has been developed. Planemo provides functionality for verifying best-practice adherence, testing, installation, and uploading of tools to the ToolShed.

Community contributions have helped the Galaxy framework and its tool suite to grow considerably. 174 developers, who have collectively produced 13,135 commits within just the past two years (63% increase since January 2016), have improved Galaxy's scalability, functionality, and usability. The project utilizes the Travis and Jenkins continuous integration (CI) services to automatically execute comprehensive test suites on each set of proposed code changes. This strategy helps prevent the introduction of bugs to the codebase and improves review time. By harnessing the open-source community and modern software development practices, we are able to release a new stable version of the Galaxy framework every four months. Current future directions include enabling data and compute federation; tighter coupling of Interactive Environments with provenance and reuse; ToolShed installation and development enhancements; continued work on collections, workflows, analysis interfaces, and history views; additional training material; improving statistical usage tracking and instrumentation; and much more. For anyone interested in getting involved with Galaxy development, we invite them to read the project's Contributing and Code of Conduct documents, review open issues, and explore the current roadmap, all which are available from the Galaxy GitHub repository (`https://github.com/galaxyproject/galaxy/`).

Acknowledgements

The Galaxy Project has grown in large part thanks to the contributions of time and effort by numerous individuals over the years. Contributing individuals include members of the Galaxy user, developer, and administrative communities and organizers of Galaxy Community Conferences. We are indebted to these helpful people. The Public Galaxy site is located at the Texas Advanced Computing Center (TACC at the University of Texas). We are extremely grateful to both TACC and CyVerse for enabling Galaxy to serve thousands of researchers worldwide. This project is supported through grant number HG006620 from the National Human Genome Research Institute, National Institutes of Health as well as grants HG005133, HG004909 and HG005542 and NSF grants DBI 0543285, 0850103, and 1661497. Additional funding is provided by Huck Institutes for the Life Sciences at Penn State and, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

Bibliography

[0] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko, "Galaxy: a platform for interactive large-scale genome analysis," *Genome Res*, vol. 15, no. 10, pp. 1451–1455, 2005.

[1] D. Blankenberg, J. Taylor, I. Schenck, J. He, Y. Zhang, M. Ghent, N. Veeraraghavan, I. Albert, W. Miller, K. D. Makova, *et al.*, "A framework for collaborative analysis of encode data: making large-scale analyses biologist-friendly," *Genome research*, vol. 17, no. 6, pp. 960–964, 2007.

[2] E. Afgan, D. Baker, M. Van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, *et al.*, "The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W3–W10, 2016.

[3] J. Yang, Y. Tanaka, M. Seay, Z. Li, J. Jin, L. X. Garmire, X. Zhu, A. Taylor, W. Li, G. Euskirchen, *et al.*, "Single cell transcriptomics reveals unanticipated features of early hematopoietic precursors," *Nucleic acids research*, vol. 45, no. 3, pp. 1281–1296, 2016.

[4] A. B. Yoo, M. A. Jette, and M. Grondona, "Slurm: Simple linux utility for resource management," in *Workshop on Job Scheduling Strategies for Parallel Processing*, pp. 44–60, Springer, 2003.

[5] D. Thain, T. Tannenbaum, and M. Livny, "Distributed computing in practice: the condor experience," *Concurrency and computation: practice and experience*, vol. 17, no. 2-4, pp. 323–356, 2005.

[6] B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A. D. Joseph, R. H. Katz, S. Shenker, and I. Stoica, "Mesos: A platform for fine-grained resource sharing in the data center.," in *NSDI*, vol. 11, pp. 22–22, 2011.

[7] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, *et al.*, "Xsede: accelerating scientific discovery," *Computing in Science & Engineering*, vol. 16, no. 5, pp. 62–74, 2014.

[8] C. A. Stewart, T. M. Cockerill, I. Foster, D. Hancock, N. Merchant, E. Skidmore, D. Stanzione, J. Taylor, S. Tuecke, G. Turner, *et al.*, "Jetstream: A self-provisoned, scalable science and engineering cloud environment," 2015.

[9] N. Goonasekera, A. Lonie, J. Taylor, and E. Afgan, "Cloudbridge: a simple cross-cloud python library," in *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale*, p. 37, ACM, 2016.

[10] D. Blankenberg, G. Von Kuster, E. Bouvier, D. Baker, E. Afgan, N. Stoler, J. Taylor, and A. Nekrutenko, "Dissemination of scientific software with galaxy toolshed," *Genome biology*, vol. 15, no. 2, p. 403, 2014.

[11] U. Paila, B. A. Chapman, R. Kirchner, and A. R. Quinlan, "Gemini: integrative exploration of genetic variation and genome annotations," *PLoS computational biology*, vol. 9, no. 7, p. e1003153, 2013.

[12] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, *et al.*, "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and environmental microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[13] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, *et al.*, "Qiime allows analysis of high-throughput community sequencing data," *Nature methods*, vol. 7, no. 5, pp. 335–336, 2010.

[14] F. Ramírez, F. Dündar, S. Diehl, B. A. Grüning, and T. Manke, "deeptools: a flexible platform for exploring deep-sequencing data," *Nucleic acids research*, vol. 42, no. W1, pp. W187–W191, 2014.

[15] F. Ramírez, D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dündar, and T. Manke, "deeptools2: a next generation web server for deep-sequencing data analysis," *Nucleic acids research*, vol. 44, no. W1, pp. W160–W165, 2016.

[16] F. Ramírez, V. Bhardwaj, L. Arrigoni, K. C. Lam, B. A. Grüning, J. Villaveces, B. Habermann, A. Akhtar, and T. Manke, "High-resolution tads reveal dna sequences underlying genome organization in flies," *Nature communications*, vol. 9, no. 1, p. 189, 2018.

[17] X. Lucas, B. A. Grüning, and S. Günther, "Chemicaltoolbox and its application on the study of the drug like and purchasable space," *Journal of Cheminformatics*, vol. 6, no. S1, p. P51, 2014.

[18] N. J. Loman and A. R. Quinlan, "Poretools: a toolkit for analyzing nanopore sequence data," *Bioinformatics*, vol. 30, no. 23, pp. 3399–3401, 2014.

[19] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "Multiqc: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, 2016.

[20] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic rna-seq quantification," *Nature biotechnology*, vol. 34, no. 5, p. 525, 2016.

[21] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon provides fast and bias-aware quantification of transcript expression," *Nature methods*, vol. 14, no. 4, p. 417, 2017.

[22] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome biology*, vol. 15, no. 12, p. 550, 2014.

[23] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "Star: ultrafast universal rna-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 2013.

[24] K. Qu, S. Garamszegi, F. Wu, H. Thorvaldsdottir, T. Liefeld, M. Ocana, D. Borges-Rivera, N. Pochet, J. T. Robinson, B. Demchak, *et al.*, "Integrative genomic analysis by interoperation of bioinformatics tools in genomespace," *nature methods*, vol. 13, no. 3, p. 245, 2016.

[25] B. A. Grüning, H. Rasche, B. Rebolledo-Jaramillo, C. Eberhard, T. Houwaart, J. Chilton, N. Coraor, R. Backofen, J. Taylor, and A. Nekrutenko, "Jupyter and galaxy: Easing entry barriers into complex data analyses for biomedical researchers," *PLoS computational biology*, vol. 13, no. 5, p. e1005425, 2017.

[26] J. Goecks, C. Eberhard, T. Too, A. Nekrutenko, and J. Taylor, "Web-based visual analysis for high-throughput genomics," *BMC genomics*, vol. 14, no. 1, p. 397, 2013.

[27] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[28] A. S. Rose and P. W. Hildebrand, "Ngl viewer: a web application for molecular visualization," *Nucleic acids research*, vol. 43, no. W1, pp. W576–W579, 2015.

[29] J. Gómez, L. J. García, G. A. Salazar, J. Villaveces, S. Gore, A. García, M. J. Martín, G. Launay, R. Alcántara, N. Del-Toro, *et al.*, "Biojs: an open source javascript framework for biological data visualization," *Bioinformatics*, vol. 29, no. 8, pp. 1103–1104, 2013.

[30] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, *et al.*, "The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud," *Nucleic acids research*, vol. 41, no. W1, pp. W557–W561, 2013.

[31] P. Amstutz, M. R. Crusoe, N. Tijanić, B. Chapman, J. Chilton, M. Heuer, A. Kartashov, D. Leehr, H. Ménager, M. Nedeljkovich, *et al.*, "Common workflow language, v1. 0," 2016.

[32] D. Blankenberg, J. E. Johnson, G. Team, J. Taylor, and A. Nekrutenko, "Wrangling galaxy's reference data," *Bioinformatics*, vol. 30, no. 13, pp. 1917–1919, 2014.

[33] J. Blomer, P. Buncic, I. Charalampidis, A. Harutyunyan, D. Larsen, and R. Meusel, "Status and future perspectives of cernvm-fs," in *Journal of Physics: Conference Series*, vol. 396, p. 052013, IOP Publishing, 2012.

[34] G. M. Kurtzer, V. Sochat, and M. W. Bauer, "Singularity: Scientific containers for mobility of compute," *PloS one*, vol. 12, no. 5, p. e0177459, 2017.

[35] F. da Veiga Leprevost, B. A. Grüning, S. Alves Aflitos, H. L. Röst, J. Uszkoreit, H. Barsnes, M. Vaudel, P. Moreno, L. Gatto, J. Weber, *et al.*, "Biocontainers: an open-source and community-driven framework for software standardization," *Bioinformatics*, vol. 33, no. 16, pp. 2580–2582, 2017.

[36] P. J. Cock, J. M. Chilton, B. Grüning, J. E. Johnson, and N. Soranzo, "Ncbi blast+ integrated into galaxy," *Gigascience*, vol. 4, no. 1, p. 39, 2015.

[37] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "Blast+: architecture and applications," *BMC bioinformatics*, vol. 10, no. 1, p. 421, 2009.

[38] B. A. Grüning, J. Fallmann, D. Yusuf, S. Will, A. Erxleben, F. Eggenhofer, T. Houwaart, B. Batut, P. Videm, A. Bagnacani, *et al.*, "The rna workbench: best practices for rna and high-throughput sequencing bioinformatics in galaxy," *Nucleic acids research*, vol. 45, no. W1, pp. W560–W566, 2017.

[39] D. Blankenberg, J. Taylor, and A. Nekrutenko, "Online resources for genomic analysis using high-throughput sequencing," *Cold Spring Harbor Protocols*, vol. 2015, no. 4, pp. pdb–top083667, 2015.

[40] L. D. Parnell, P. Lindenbaum, K. Shameer, G. M. Dall'Olio, D. C. Swan, L. J. Jensen, S. J. Cockell, B. S. Pedersen, M. E. Mangan, C. A. Miller, *et al.*, "Biostar: an online question & answer resource for the bioinformatics community," *PLoS computational biology*, vol. 7, no. 10, p. e1002216, 2011.

Galactic Circos: User-friendly Circos Plots within the Galaxy platform

Helena Rasche [1,*], Saskia Hiltemann[1,*]

* Helena Rasche and Saskia Hiltemann contributed equally to this work.

Published in: *GigaScience*, Volume 9, Issue 6, June 2020, giaa065
DOI: https://doi.org/10.1093/gigascience/giaa065

1. Bioinformatics Group, Department of Computer Science, University of Freiburg, 79110 Freiburg im Breisgau, Germany
2. Erasmus Medical Center, Clinical Bioinformatics Group, Department of Pathology, Rotterdam, The Netherlands.

Abstract

Background: Circos is a popular, highly flexible software package for the circular visualization of complex datasets. While especially popular in the field of genomic analysis, Circos enables interactive graphing of any analytical data, including alternative scientific domain data and non-scientific data. This high degree of flexibility also comes with a high degree of complexity, which may present an obstacle for researchers not trained in programming or the UNIX command line. The Galaxy platform provides a user-friendly browser-based graphical interface incorporating a broad range of *"wrapped"* command line tools to facilitate accessibility.

Findings: We have developed a Galaxy wrapper for Circos, thus combining the power of Circos with the accessibility and ease of use of the Galaxy platform. The combination substantially simplifies the specification and configuration of Circos plots for end users while retaining the power to produce publication-quality visualizations of complex multidimensional datasets.

Conclusions: Galactic Circos enables the creation of publication-ready Circos plots using only a web browser, via the Galaxy platform. Users may download the full set of Circos configuration files of their plots for further manual customization. This version of Circos is available as an open-source installable application from the Galaxy ToolShed, with its use clarified in a training manual hosted by the Galaxy Training Network.

Keywords: Genomics; Visualisation; Galaxy; Circos; UI/UX

Findings

Background

The Circos visualization tool [0] is widely used in the biological scientific community and is especially popular for use in scientific publications. Circos has >4000 citations, and its plots have appeared on the cover of several leading scientific journals [1]. Its popularity is due in large part to its great flexibility; Circos offers a wide range of visualization options, and all aspects of a Circos plot may be customized to the user's needs. While originally created for the visualization of genomic data, Circos makes no a priori assumptions about the format and domain of the input data; this is illustrated by the fact that it has been used for a wide range of applications, from genomics research to visualizations of car sales, urban planning, and presidential debates [2].

With Circos's great flexibility also comes a high degree of complexity and a significant learning curve, and as a result its use is often limited to expert users who are experienced with programming and the UNIX command line.

The Galaxy platform [3] aims to provide a user-friendly interface to command line tools and empower domain experts to run powerful analysis and visualization tools without the need for any programming experience. Galaxy offers a wide range of tools for a variety of applications domains and is widely used in the biological scientific community (8900+ citations, 7500+ tools [4, 5]). Galaxy also automates the installation of tools and all their dependencies, removing another hurdle for its use by research scientists.

Our tool combines the power of Circos with the user-friendliness of the Galaxy interface to greatly increase the accessibility of the tool and simplify the creation of publication-ready plots for scientific data.

Previously, custom Circos Galaxy plotter tools have been written [6]; however, these tools are not generic, but are tailored specifically to the use-case at hand. This means that a new Galaxy tool has to be created whenever a new plot type is needed. Galactic Circos aims to be a generic tool capable of creating any Circos plot regardless of data domain.

Results

The Galactic Circos tool changes the way users must specify the configuration of a Circos plot. Instead of writing a number of configuration files, users now only need to select the various plot options from a web interface, and datasets from their analysis history (Figure 1.3). Because Circos plot specifications can be quite complex, the tool interface is subdivided into several collapsible sections, each corresponding to a different Circos configuration option in order to increase the usability of the tool. Parameters are preconfigured with sensible default values so that basic plots can be generated with minimal configuration.

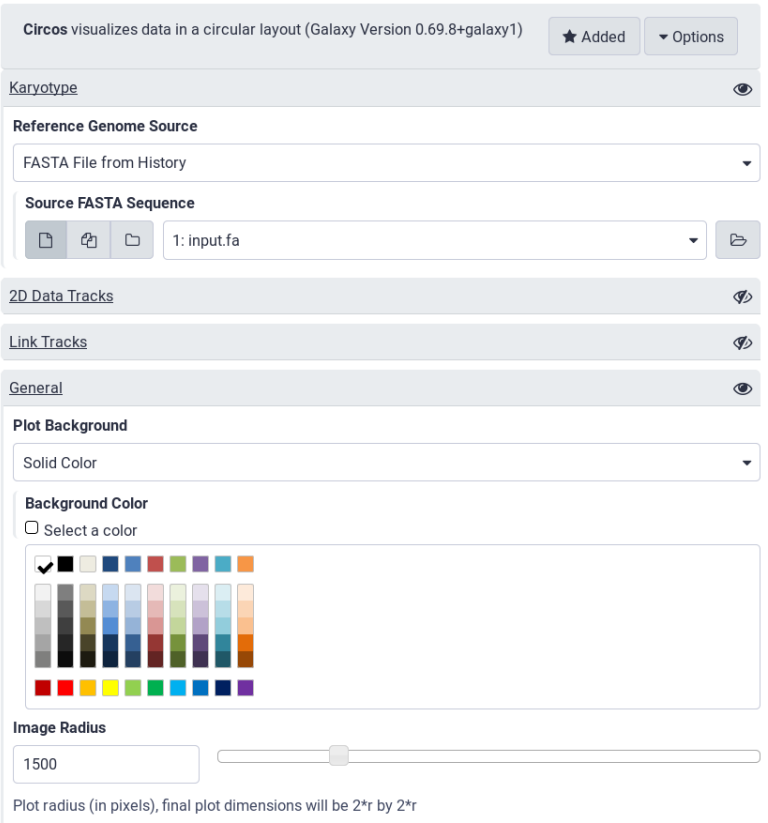


Figure 1.3: The Galaxy tool interface to Circos. Each collapsed section hides a wealth of configuration options available to users. The web-based interface is significantly more accessible than the command line version.

We demonstrate the utility of the Galactic Circos tool by recreating one of the more advanced examples from the Circos online tutorials, the microbial genome lesson [7] (Figure 1.4). This displays multiple tracks of different types (text, histogram, tiles), has a customized ideogram, and uses rules for colouring data points dependent on their value.

In a second example (Figure 1.5), we replicate within Galaxy the cover image of the *Nature* issue [8] dedicated

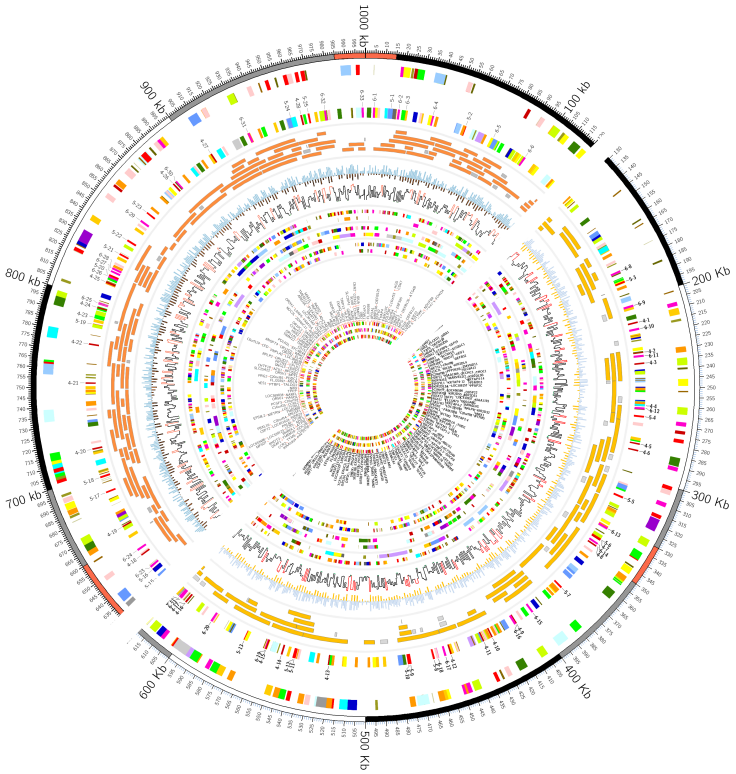


Figure 1.4: Here we reproduce one of the more complex tutorials from the Circos documentation. The upper left half of the image is produced by the configuration provided by the Circos tutorial, while the bottom-right half is produced completely in Galaxy. While some options used in the original tutorial cannot be directly used (e.g. unrestricted perl code), they can be recreated equivalently in the tool interface. Some options in the tool interface are likewise restricted; Galactic Circos offers a color picker with a limited palette, which accounts for the differences in colour. However, our tool offers the ability to download the full Circos configuration folder, allowing advanced users to configure the colour (or other) parameters manually and rebuild the image locally. See `https://usegalaxy.eu/u/helena-rasche/h/circos-microbe-tutorial`.

to the ENCODE project [9]. This cover featured a Circos plot and is also available as part of the official Circos tutorials [10].

These 2 examples showcase a variety of different track types (histograms, scatterplot, highlights, tiles, text) and configurations (ticks, rules, ideogram customizations) to illustrate the feature-completeness of Galactic Circos.

Workflow summary

Visualizations in the Galaxy framework are usually implemented as interactive JavaScript components, but these plots cannot be created automatically in workflows. Individual plotting tools exist as Galaxy tools; however, these are less common and generally less flexible because tool authors must make a trade-off between development time and feature support. We put significant time into the development in order to make

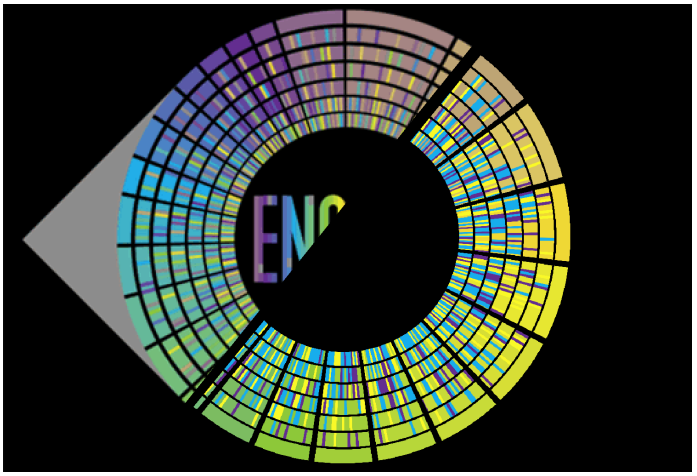


Figure 1.5: *Nature* cover for the ENCODE project in September 2012, reproduced by Galactic Circos. The image is not a split image owing to copyright restrictions on the original cover image. Comparison can be made against the Circos tutorial [10]. https://usegalaxy.eu/u/helena-rasche/h/circos-encode-nature-cover

an extremely generic tool, enabling researchers to use the Galactic Circos tool in their workflows, based on previous experiences building single-purpose Circos plotting tools (e.g., as in Figure 1.6). This enables creation of human-readable summaries of large analysis workflows, similar to the non-genomics–focused iReport [11]. Galactic Circos was born from precisely this use-case and therefore aims to enable reducing complex analysis pipeline outputs, such as the workflows required in cancer genomics, allowing bioinformaticians to produce a single image summarizing all of their relevant outputs in an easily digestible manner.

Supporting Tools

Circos requires input datasets to adhere to a specific and custom file format. To facilitate the conversion of data to this custom Circos format, we have developed several supporting Galaxy tools for conversion. These tools allow users to convert their datasets from a variety of common genomics formats such as (big)Wig files, interval files, and MAF/Stockholm alignments. Furthermore, the existing Galaxy ecosystem provides a wide array of tabular data manipulation tools that can be leveraged to transform any tabular or text files into the format accepted by Circos.

To demonstrate the utility of these supporting tools, we show a real-world example of a plot using common genomics datasets. This example is a recreation of a plot in a published paper demonstrating chromothripsis in the VCaP prostate cancer cell line [12]. The input datasets originate from a variety of sources, including a structural variants file (converted to Circos links track), copy number and B-allele frequency track obtained from Affymetrix single-nucleotide-polymorphism (SN)P array data, and a SNP density track generated from a VCF file. Using a combination of the supporting tools included in the Galactic Circos package and the generic file manipulation tools present in Galaxy, we were able to convert these various datasets to Circos-compatible

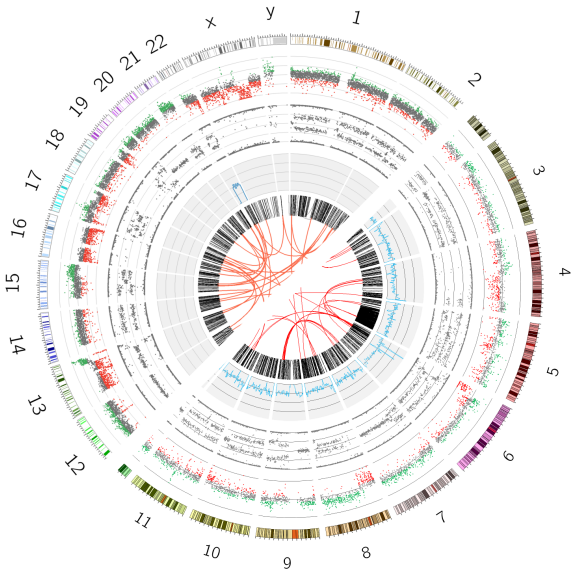


Figure 1.6: In the top panel, comparison of the output of a custom-written Circos plot with hard-coded configuration (upper left half) to the output created using the Galactic Circos tool (lower right half). While the input data originated from a range of standard and non-standard genomic file formats, conversion to Circos-formatted files was possible using the plethora of file manipulation tools already integrated into Galaxy and the set of supporting conversion tools included in the Galactic Circos package. In the bottom panel we produce Circos plots per chromosome, leveraging Galaxy's ability to map a tool execution across a collection of input datasets, in this case each karyotype in a separate input file. The images are reduced and placed together in a montage using further Galaxy tools. `https://usegalaxy.eu/u/helena-rasche/h/circos-cancer-genomics--chromothripsis`, `https://usegalaxy.eu/u/helena-rasche/h/circos-multiplot`.

formats without leaving Galaxy, and reproduced the Circos plot from the publication (Figure 1.6).

Once data have been reformatted for Circos, they can either be used immediately or be further processed. Circos includes a tool suite for post-processing and downsampling of data, which can improve plot clarity and processing speed. We additionally included a number of these post-processing tools into Galaxy, notably the link-bundling and binning tools used in Figure 1.7.

Finally, while Circos is widely used for the visualization of genomic data, and many of the parameter names have a distinctly biological feel to them, the tool does not impose any restrictions on the type of input data, and is capable of displaying non-biological data just as easily [2]. To show that our tool retains this degree of flexibility, we recreated the presidential debate plot included in the Circos tutorials, which in turn was based on a plot which appeared in the New York Times article [14]. A plot comparison can be seen in Figure 1.8.

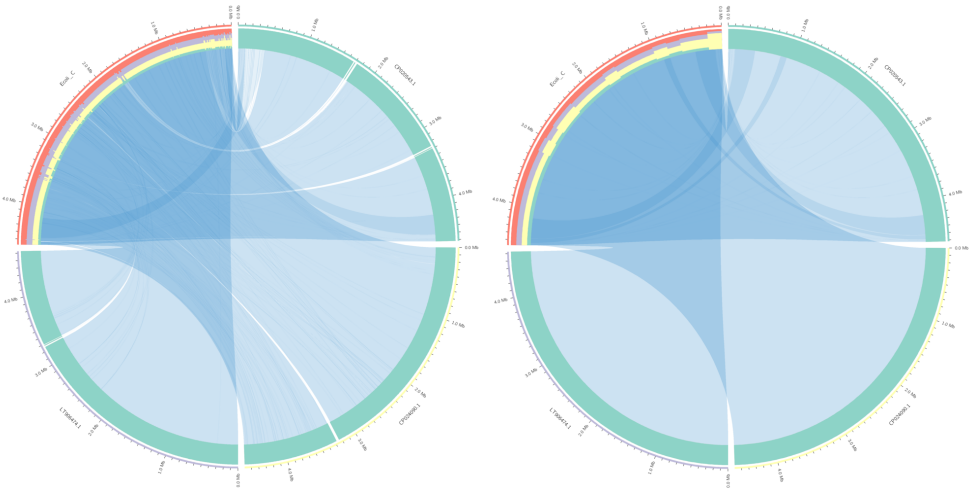


Figure 1.7: These 2 plots show the link-binning and bundling scripts used with different thresholds. The inner link track was generated directly from a MAF file output by LastZ [13]. This file was processed by Circos's bundling tool in Galaxy to decrease the number of links, a process usually done to decrease visual noise and increase efficiency. The outer track demonstrates the link-binning script, which generates a histogram, in this case from the number of links to that position in the genomic region.

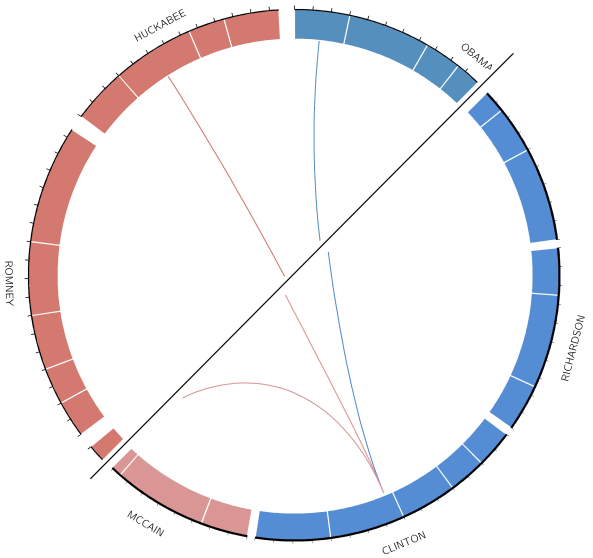


Figure 1.8: This figure compares the Circos plot from the official tutorial (upper left half), to the output created using the Galactic Circos tool (lower right half). Each link represents a candidate speaking the last name of another candidate. The length of each circle segment is proportional to the total number of words spoken by the candidate during the debates. `https://usegalaxy.eu/u/saskia/h/circos-politics-plot`.

Lessons learned and limitations

Given the great flexibility and configurability of the Circos tool, our Galaxy wrapper is, to our knowledge, one of the most complex Galaxy tools. Development of this wrapper took significant time and resources, and in places took us to the edges of what is possible in Galaxy. In this section we describe some of the lessons learned and tips for wrapping tools of this complexity.

Security

This wrapper exposes ~95% of what is possible with Circos. We intentionally excluded the last ~5% of features because we could not safely implement them. These features would require allowing free-text user input and unrestricted Perl code, which can pose a potential security risk. We believed that we could not, within a reasonable period of development time, implement sufficient sanitization of all possible user inputs. Instead, we provide an option for the tool to output the full set of configuration files required to recreate the plot, which the user can use as a starting point for manual adaptation locally. There are ongoing efforts within the Galaxy community to perform computations with increasingly untrusted user input, and we hope that the Galaxy community will push this even further in the future and make it the default, rather than requiring special configuration and knowledge from system administrators. This would enable us to add a free-text field within the Circos tool, and users could provide custom configuration freely and without risk to the administrator.

Visualization vs tool

We made the initial choice to build Galactic Circos as a tool, not a visualization, given the long compilation times of plots and our desire to build a workflow-compatible tool because this was not possible in Galaxy at that time. In the future, we might want to explore the possibilities of a more dynamic visual interface, using a visualization plugin in Galaxy. We would have complete freedom to build in more interactivity and custom components (e.g., visual preview for Brewer scale selection) as needed.

Macros

Macros proved incredibly helpful in wrangling the complexity of this tool by allowing us to define reusable components and avoid code duplication. Galaxy wrappers allow for the definition of "macros"; these are bits of code defined in a file outside the main wrapper and can be reused at multiple points in the tool. Unfortunately, the extent to which this tool relies on macros also makes the tool more complex from a development point of view, with the code spread out over a large number of files. However, the benefits here outweigh the drawbacks.

Collapsible sections

The section feature in Galaxy permits grouping related options together in the user interface. This avoids overwhelming the user with the enormous array of available parameters but rather groups these logically and only shows those subsets requested by the user. Unfortunately, these sections re-collapse themselves during tool rerun and are not marked when their children contain modifications from the defaults. If this was changed, users could more easily recall what they did in the previous tool run because all edited sections would be expanded, or marked by default.

Color
The built in color selector provides a small palette of colors. While it is a good thing to prevent users from making plots with hard-to-see or unpleasant colors, it also substantially limits more advanced users. The addition of an advanced color picker would be welcome for Circos users. Likewise we used a select box for Brewer palette, which feels suboptimal compared to a component that could include a preview of that palette and would be much more user-friendly.

Methods

Implementation

The execution of the tool leverages Galaxy's ability to write templated files directly to disk with configuration from the tool form, and then running Circos directly on these templated configuration files.

Installation of the Circos tool and its dependencies is handled by the Galaxy platform, which supports different dependency management frameworks, including Conda and Containers. All dependencies including Circos itself are available from the Bioconda Conda channel [15] and available as a virtualised container (rkt, Docker, Singularity). The version of the Galaxy Circos tool being reported on here uses Circos version 0.69.8.

File Format Converters

To facilitate the interoperability with upstream tools and workflows, we provide a set of file format converters, in addition to many tools already available in Galaxy, which together provide for conversion of a range of common data format standards (e.g. VCF, MAF/Stockholm, BED/GFF3, BigWig). These tools produce files that are ready to be used as input to the Galaxy Circos tool. Additionally the applicable subset of circos-utils were included into Galaxy for Circos-friendly tools for data reshaping.

Circos Configuration Export

While Galactic Circos aims to offer the full range of Circos functionality, some manual customization of the Circos plot configuration files may still be desired. To this end, our tool also outputs the full set of configuration files needed to recreated the plot on the command line, and thus allow easy access to any features not exposed in the Galaxy wrapper.

Training Materials

Our tool greatly simplifies the creation of Circos plots, but the large number of options offered by the Circos tool necessitates good documentation and explanation to optimize their utility for end-users. Circos offers a collection of tutorials that are designed to familiarize users with the various features of Circos [16]. In a similar

fashion, we have created a set of Galaxy tutorials aimed to educate users in the use of Circos within Galaxy. These tutorials are available from the Galaxy training materials website [17].

Reproducible and Reusable Plots

To enable readers to examine the complete parameters settings used and recreate the example plots given here, Galaxy histories for all the figures shown in this work have been made publicly available from the European Galaxy server (see Availability section).

Future Work

While we have aimed to make our tool as feature-complete as possible, some of Circos's functionality is not currently exposed in the Galaxy tool. We intend to extend our tool to include these features, including but not limited to support for scaling subsections of the plots, and generation of HTML image maps.

Availability of Source Code and Requirements

- Project name: Galactic Circos
- Github repository: `https://github.com/galaxyproject/tools-iuc/tree/master/tools/circos`
- Tool Shed repository: `https://toolshed.g2.bx.psu.edu/view/iuc/circos`
- Training Manual: `https://training.galaxyproject.org/training-material/topics/visualisation/tutorials/circos/tutorial.html`
- Operating system(s): Unix (Platform independent with Docker)
- Other requirements: Galaxy version 18.01 or higher
- License: GNU GPL

The Circos example plots presented in this work are available as Galaxy histories:

- Galaxy history for Figure 2: `https://usegalaxy.eu:/u/helena-rasche/h/circos-microbe-tutorial`
- Galaxy history for Figure 3: `https://usegalaxy.eu/u/helena-rasche/h/circos-encode-nature-cover`
- Galaxy history for Figure 4: `https://usegalaxy.eu/u/helena-rasche/h/circos-cancer-genomics--chromothripsis`

Galaxy Resources

- Galaxy Home Page: https://galaxyproject.org
- Galaxy Tutorials: https://training.galaxyproject.org
- How to install Galaxy: https://getgalaxy.org
- How to install tools: https://galaxyproject.org/admin/tools/add-tool-from-toolshed-tutorial/
- Full Administrative resources: https://docs.galaxyproject.org
- Galaxy Help Forum: https://help.galaxyproject.org
- Connect with the Galaxy Community on Gitter Chat: https://gitter.im/galaxyproject/Lobby/
- Public Galaxy servers that include Circos: usegalaxy.eu, usegalaxy.org, usegalaxy.org.au (see Galactic Circos tutorial for full up-to-date list).

Availability of Supporting Data and Materials

The data presented here to illustrate our application was obtained from previous publications, and has been collected and made available from Zenodo.

Additional supporting data are available from the GigaScience GigaDB database.

Declarations

Abbreviations

BED: Browser Extensible Data; ENCODE: Encyclopedia of DNA Elements; GFF: general feature format; HTML: HyperText Markup Language; MAF: multiple alignment format; SNP: single-nucleotide polymorphism; VCF: variant call format

Competing Interests

The authors declare that they have no competing interests.

Funding

This project was made possible with the support of the Albert Ludwig University of Freiburg and German Federal Ministry of Education and Research (031 L0101C de.NBI-epi).

Funding for open access charge: German Federal Ministry of Education and Research.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement 825775

Author's Contributions

HR and SH contributed to the tool development, documentation, and writing of the manuscript.

Acknowledgements

The authors would like to thank the Galaxy community for their help in reviewing, testing, and validating the tools presented here.

Bibliography

[0] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, "Circos: An information aesthetic for comparative genomics," *Genome Research*, vol. 19, pp. 1639–1645, jun 2009.

[1] "Scientific literature images created with circos." http://circos.ca/images/published/.

[2] "Using circos to visualize non-genomic (general) data." http://circos.ca/intro/general_data/.

[3] E. Afgan, D. Baker, B. Batut, M. Van Den Beek, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, B. A. Grüning, *et al.*, "The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update," *Nucleic acids research*, vol. 46, no. W1, pp. W537–W544, 2018.

[4] "Zotero list of citations of the galaxy project." https://www.zotero.org/groups/1732893/galaxy.

[5] "Galaxy tool shed." https://toolshed.g2.bx.psu.edu/.

[6] S. Hiltemann, H. Mei, M. de Hollander, I. Palli, P. van der Spek, G. Jenster, and A. Stubbs, "Cgtag: complete genomics toolkit and annotation in a cloud-based galaxy," *GigaScience*, vol. 3, no. 1, p. 1, 2014.

[7] "Circos microbial genome lesson." http://www.circos.ca/documentation/tutorials/recipes/microbial_genomes/images.

[8] "Nature encode cover (volume 489 issue 7414)." https://www.nature.com/nature/volumes/489/issues/7414.

[9] E. P. Consortium *et al.*, "The encode (encyclopedia of dna elements) project," *Science*, vol. 306, no. 5696, pp. 636–640, 2004.

[10] "Circos encode nature cover image lesson." http://www.circos.ca/documentation/tutorials/recipes/nature_cover_encode/.

[11] S. Hiltemann, Y. Hoogstrate, P. Van der Spek, G. Jenster, and A. Stubbs, "ireport: a generalised galaxy solution for integrated experimental reporting," *GigaScience*, vol. 3, no. 1, p. 19, 2014.

[12] I. T. Alves, S. Hiltemann, T. Hartjes, P. van der Spek, A. Stubbs, J. Trapman, and G. Jenster, "Gene fusions by chromothripsis of chromosome 5q in the vcap prostate cancer cell line," *Human genetics*, vol. 132, no. 6, pp. 709–713, 2013.

[13] A.-M. Rahmani, P. Liljeberg, J. Plosila, and H. Tenhunen, "Lastz: An ultra optimized 3d networks-on-chip architecture," in *2011 14th Euromicro Conference on Digital System Design*, pp. 173–180, IEEE, 2011.

[14] "Naming names - interactive graphic - nytimes.com." http://archive.nytimes.com/www.nytimes.com/interactive/2007/12/15/us/politics/DI

[15] B. Grüning, R. Dale, A. Sjödin, B. A. Chapman, J. Rowe, C. H. Tomkins-Tinch, R. Valieris, and J. Köster, "Bioconda: sustainable and comprehensive software distribution for the life sciences," *Nature methods*, vol. 15, no. 7, p. 475, 2018.

[16] "Circos tutorials." http://circos.ca/tutorials/lessons/.

[17] B. Batut, S. Hiltemann, A. Bagnacani, D. Baker, V. Bhardwaj, C. Blank, A. Bretaudeau, L. Brillet-Guéguen, M. Čech, J. Chilton, D. Clements, O. Doppelt-Azeroual, A. Erxleben, M. A. Freeberg, S. Gladman, Y. Hoogstrate, H.-R. Hotz, T. Houwaart, P. Jagtap, D. Larivière, G. L. Corguillé, T. Manke, F. Mareuil, F. Ramírez, D. Ryan, F. C. Sigloch, N. Soranzo, J. Wolff, P. Videm, M. Wolfien, A. Wubuli, D. Yusuf, J. Taylor, R. Backofen, A. Nekrutenko, and B. Grüning, "Community-driven data analysis training for biology," *Cell Systems*, vol. 6, pp. 752–758.e1, jun 2018.

iReport: A generalised Galaxy solution for integrated experimental reporting

Saskia Hiltemann[1,2], Youri Hoogstrate[1,2], Peter van der Spek[1], Guido Jenster[2], Andrew Stubbs[1]

Published in: *GigaScience*, Volume 3, Issue 1, December 2014, 2047-217X-3-19
DOI: https://doi.org/10.1186/2047-217X-3-19

1. Department of Bioinformatics, Erasmus Medical Center, Rotterdam, The Netherlands.
2. Department of Urology, Erasmus Medical Center, Rotterdam, The Netherlands.

Abstract

Background: Galaxy offers a number of visualisation options with components, such as Trackster, Circster and Galaxy Charts, but currently lacks the ability to easily combine outputs from different tools into a single view or report. A number of tools produce HTML reports as output in order to combine the various output files from a single tool; however, this requires programming and knowledge of HTML, and the reports must be custom-made for each new tool.
Findings: We have developed a generic and flexible reporting tool for Galaxy, iReport, that allows users to create interactive HTML reports directly from the Galaxy UI, with the ability to combine an arbitrary number of outputs from any number of different tools. Content can be organised into different tabs, and interactivity can be added to components. To demonstrate the capability of iReport we provide two publicly available examples, the first is an iReport explaining about iReports, created for, and using content from the recent Galaxy Community Conference 2014. The second is a genetic report based on a trio analysis to determine candidate pathogenic variants which uses our previously developed Galaxy toolset for whole-genome NGS analysis, CGtag. These reports may be adapted for outputs from any sequencing platform and any results, such as omics

data, non-high throughput results and clinical variables.
Conclusions: iReport provides a secure, collaborative, and flexible web-based reporting system that is compatible with Galaxy (and non-Galaxy) generated content. We demonstrate its value with a real-life example of reporting genetic trio-analysis.

Findings

Structured reporting and documentation of experimental outcome is required for the successful transfer of knowledge from the research scientist to their peers and to the broader academic community.

Galaxy is a platform that aims to provide complex bioinformatics services and tools in an easy-to-use web-based graphical user interface [0, 1, 2]. The output from these tools can be displayed using built-in Galaxy visualisation applications [3], via specialised visuals implemented as a component in the workflow deployed in Galaxy [4] or by downloading the results and visualising the output with applications external to Galaxy (e.g., Excel, TIBCO spotfire, R, spreadsheet programs, *etc*).

Galaxy has the capacity to track the provenance of the source data, the workflow, as well as the workflow components used to analyse the data. Currently users can share their workflow and results within Galaxy, but do not have access to a simple method to summarise results from multiple tools and/or workflows in an integrated report. To address this issue we have developed iReport, an integrated reporting application that provides users with a flexible means to produce dynamic HTML reports which can be shared with other Galaxy users or downloaded to disk.

Systems used by end-users to deliver graphical output range from open-source applications, such as *Ad Hoc* reports [5], Google charts (and docs) [6] and OpenOffice [7], to commercial applications such as Microsoft Office. Indeed scientific reporting applications both open-source (Bioconductor [8], Circos [9][10]) and commercial software (e.g., Omniviz [11], Partek [12]) include a multitude of visualisation capabilities with a focus on data reporting and presentation of data in the context of the experimental design and with associated meta-data. There are some applications, like TIBCO spotfire [13], which are capable of integrating results from multiple sources including associated text and meta-data and other applications which serve as an electronic lab note book (e.g., IDBS [14]). Additionally there have been many products developed to address the selection and reporting of variants for pathogenic variant selection including the workflow to identify those variants (e.g., Gensight [15], Cartagenia [16], Clincial Genomicist [17]). For data generated in R, dynamic reporting packages such as KnitR [18], Sweave [19] and R-Markdown [20],

allow for the integration of data-generating code within the report specification itself. Similar systems exist for other programming languages, for example Tangle [21] (JavaScript), Active Markdown [22](CoffeeScript) or IPython Notebooks [23] (Python). These are very versatile tools, but require programming knowledge to use effectively. iReport offers an open-source application for both Galaxy and non-Galaxy produced results allowing for the generation of customized integrated reports for any type of project or workflow. The advantage to Galaxy users is that the output from any application can be included into any report, and that a report template can be reused for other projects. Also, the report may be securely shared with one or many users of that Galaxy instance or made publicly available. iReports can be completely configured from the tool web interface, and requires no programming or knowledge of the underlying system.

We demonstrate iReport's utility through an example where a genetic report is generated from outputs of an existing Galaxy next-generation sequencing (NGS) toolkit, CGtag [4]. iReport can also be used as an electronic lab notebook by creating an iReport which links out to various other iReports containing different analysis reports from various samples. It can be also be coupled to output from other Galaxy instances, for example output generated by specialized Galaxy instances such as Confero [24], ORIONE [25], and Galaxy-P [26].

Functionality

iReport dynamically generates HTML, and employs JavaScript and jQuery to create interactive components, such as searchable, sortable, paginated tables and zoomable images. iReport is ideally suited to use as the final step in a workflow; the pipeline developer configures the report once and end-users are then presented with a templated report each time users run the workflow, while only needing to provide the input files for the pipeline [27]. iReport can also be used directly by end-users as a means of easily sharing their results with other Galaxy users, or the public via Galaxy's native sharing capabilities.

The generic reporting functionality and usage of iReport is outlined below using an example iReport created for the recent Galaxy Community Conference, which is also available for viewing online [28]. It is followed by an example of a genetic report that can be used for trio analysis, which can easily be modified for any trio reporting or extended to quartets or larger families, also available from our demo galaxy [29].

iReport Structure

iReport produced a report webpage consisting of one or more subpages with one or multiple elements included on each subpage. The primary output of iReport is:

1. A cover page
   (a) Title of the report
   (b) Cover image
2. Main report page consisting of a set of tabs. Each tab consisting of one or more *content items*. Each content item can be one of the following types:
   (a) Text
   (b) Images
   (c) Tables
   (d) PDF Files
   (e) Links

An iReport tutorial has been developed to demonstrate and explain the functionality of iReport, and is available as a shared history from the CTMM-TraIT public Galaxy instance [28]. The following sections describe each of the components of iReport in more detail.

Cover Page

The cover page consists of a user-specified title and a cover image. The cover image parameter is optional and when the field is left blank a default image is used (Figure 1.9). By clicking on the image, or the link above it, the user can access the main report page. There is also a link to download the entire iReport webpage, including all dependency files, for storing or viewing on different systems.

Main Report Page

An arbitrary number of tabs may be added via a repeat parameter. Each tab can be labelled with a name specified by the user. An arbitrary number of *content items* may then be added to each tab in a repeat parameter. A type must be specified for each content item (e.g., text, image, table *etc.*), as well as several other parameters depending on the type chosen (Figure 1.10). Layout is mostly left up to the browser, but users can explicitly add a line-break after each item to force items to appear underneath each other.

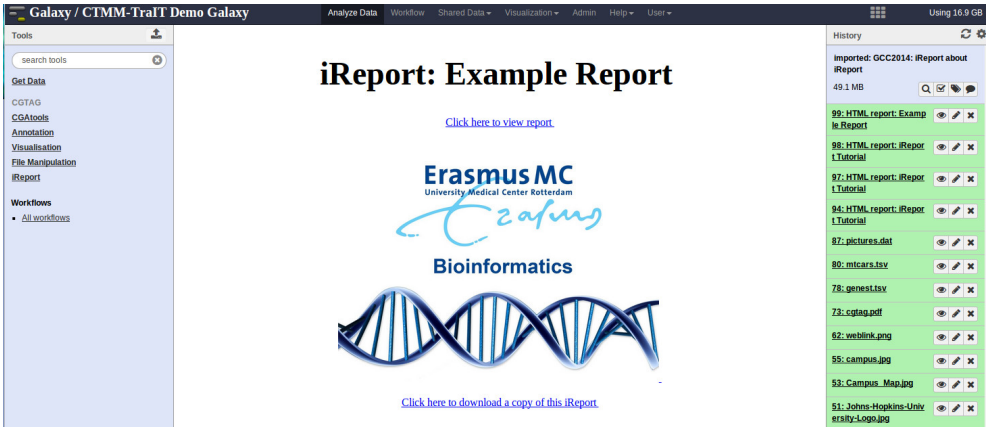


Figure 1.9: Example cover page.Example of a cover page with title *Example Report* and the default cover image. A link to download the entire iReport web page is also provided.

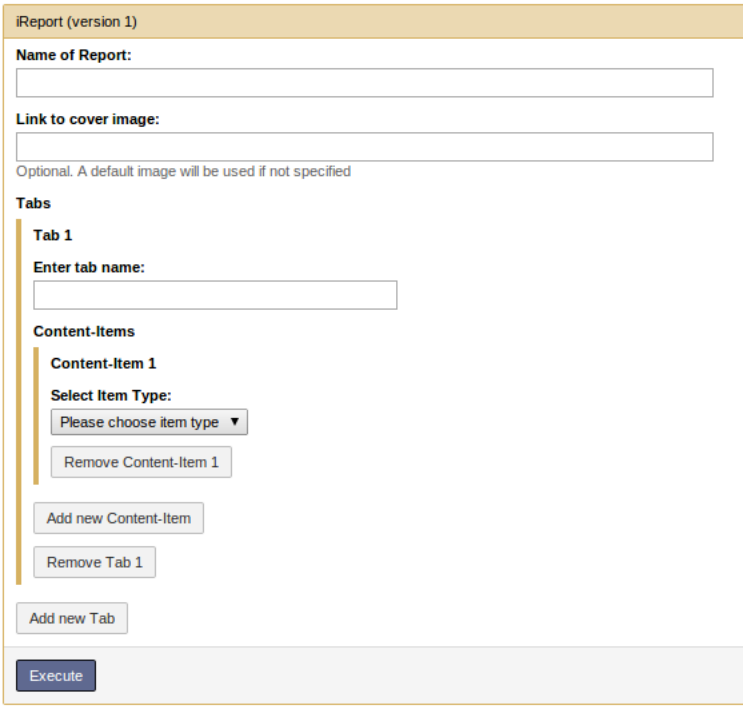


Figure 1.10: iReport tool wrapper. iReport's tool interface. Minimally a report title and at least 1 tab with 1 content item need to be specified.

Content item: Text Field Text can be entered in a text field in the tool interface, for example to create an introduction paragraph and to give a description of the items on the page. Text is printed verbatim, although a small number of HTML tags are allowed in order to give the user some control over formatting (e.g., `b, i, em, strong, h1-h6` tags). Text files can also be specified, and the contents of the file will be printed to the screen verbatim.

Content item: Images Many tools produce images as output, which can also be displayed by iReport. Users specify the image file from their Galaxy history, and the desired image size. For images that have been scaled down, an optional jQuery zoom-on-mouseover effect may be added (Figure 1.11)[30]. Currently supported image formats are JPG, PNG and SVG.

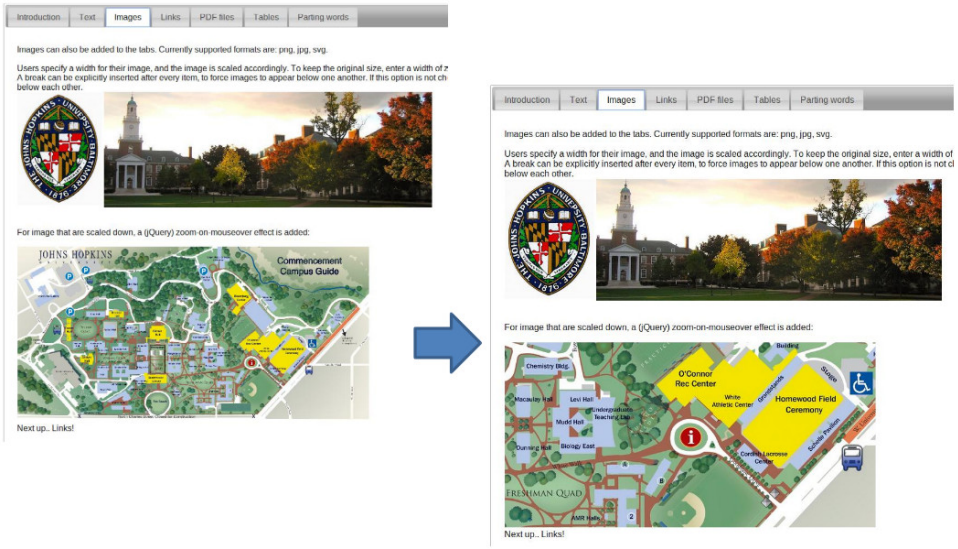


Figure 1.11: Zoom effect.Images that have been scaled down can optionally be enhanced with a jQuery zoom-on-mouseover effect. In this example, the bottom image has this effect added, and when the user moves their mouse over the image, a zoomed version of that area of the image is shown.

Content item: Tables iReport can also display tables. The input must be a tab-delimited file from the users' Galaxy history, and the first nonempty line not starting with a hash symbol (#) is assumed to contain the column headers. The jQuery library *DataTables* [31] is used to create tables which can be searchable, sortable and paginated, if requested by the user. There is an option to create hyperlinks within the columns of a table by providing a column number, a URL prefix and a URL suffix. This is illustrated in Figure 1.12, where the first column contains

gene names and by including the GeneCards [32, 33] URL prefix `http://www.genecards.org/cgi-bin/carddisp.pl?gene=`. This generates a hyperlink to the corresponding GeneCards entry for every item in the column in the table.

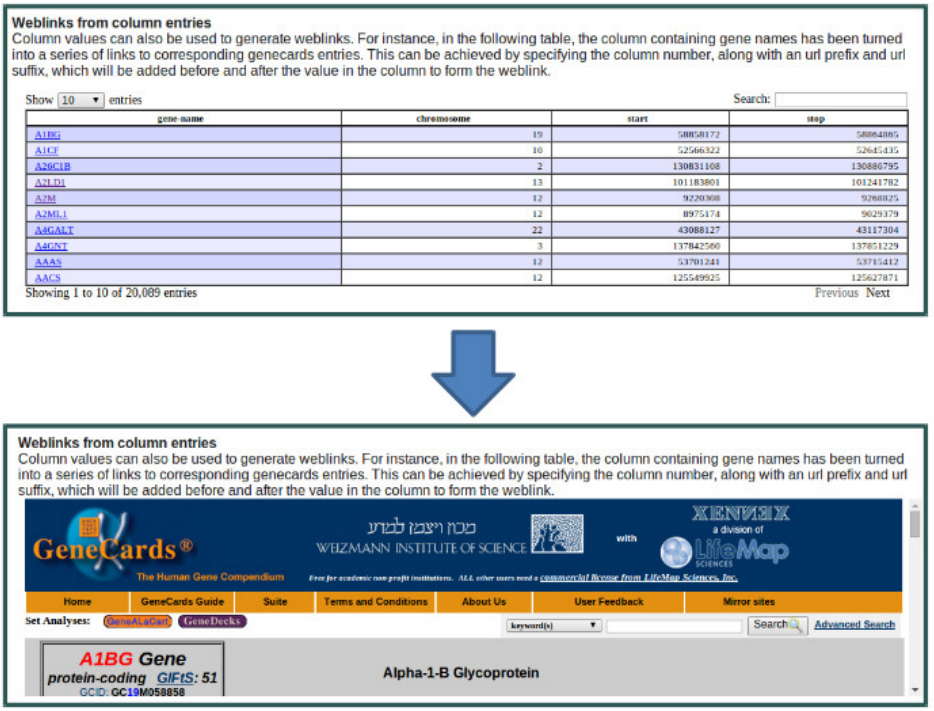


Figure 1.12: Weblinks from table columns A series of web links can be created within a table by specifying a prefix and suffix to be placed before and after each column entry.

Content item: PDF Files This is one of the simplest content items. The user provides a PDF file from the Galaxy history, which will be embedded in the page. If the browser does not have the necessary plug-ins installed, a download link for the file will be generated instead (Figure 1.13).

Content item: Links Users can create links to web locations by specifying a URL and a link text. Links to datasets in the history can also be created here by specifying a dataset and a link text. Several tools create archives of files as output (for example a zip file containing the plots for each chromosome). Links to all files contained in an archive can also be created, and will be named with the file names (excluding file extension). Currently the supported archive formats are zip, bz2, tar,

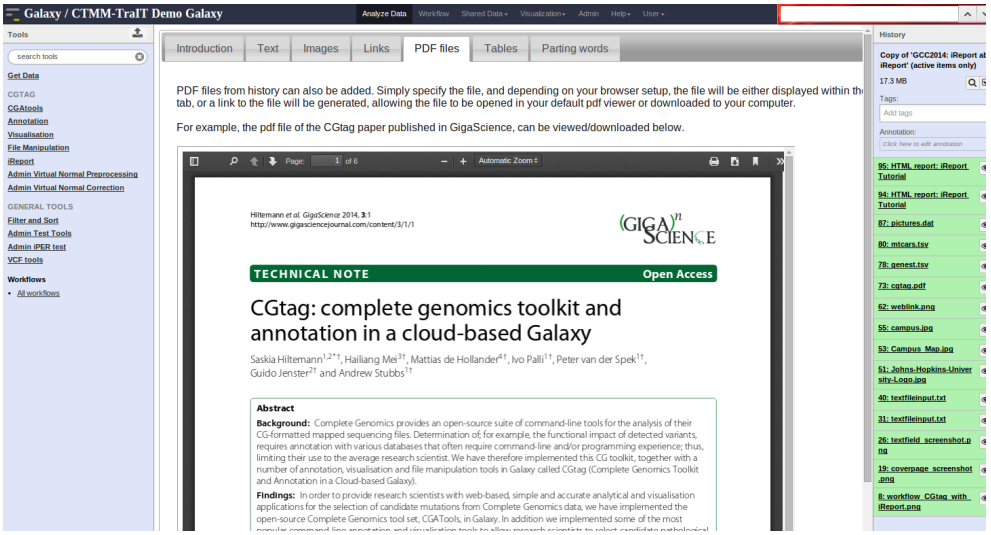


Figure 1.13: Embedded PDF files. iReports can also display PDF files. For browsers without PDF plug-in, a download link to the file will be created instead.

gz and tar.gz. An example can be seen in Figure 1.14, where an archive with images was used as input and a series of links to each contained file was created. An option to create a link to an iReport is also present. This allows users to create a kind of electronic lab notebook, by creating an overview of all their samples and linking to one or more iReports for each sample.

Genetic report for a trio of HapMap individuals

Accurate, reproducible and traceable reporting is an essential requirement to the evaluation of the genetic outcome from any assay [34], including those variations predicted from NGS analysis. Since iReport is capable of including many formats, we have used the outcome from a trio analysis generated from the Complete Genomics [35] NGS platform to demonstrate its utility in representing these data in a user-defined format, which contains the provenance of the underlying analysis. In this example we use a trio of individuals sequenced in the International HapMap Project [36][37], to demonstrate how to select protein affecting candidate variants based on a recessive genetic model. All data in this example is freely available for download from the Complete Genomics website [38].

This example iReport has one tab devoted to explaining the protocol used (Figure 1.15B), one tab with circos plots and an explanation of the family structure (Figure 1.15D), and one tab with tables

Figure 1.14: Links to all files in an archive.Given an archive of files, iReport can create a series of links to all files contained in the archive. Link texts are the filenames (without file extension).

containing the candidate pathogenic variants determined by the protocol based on a recessive model for selection. This iReport is also available as a published history on the TraIT-CTMM public Galaxy [39].

Conclusions

iReport is a easy-to-use, flexible tool for generating traceable, standardized reports which are easily shared between users within and across platforms. We have demonstrated that iReport is capable of creating a customised genetics report from results generated within Galaxy and may be shared with collaborators on the same platform, or with the public. Additionally, data or results generated externally can be uploaded into Galaxy and can also be used by iReport. These reports are generated as web pages and may be downloaded in their entirety to be easily shared across systems.

The genetics report presented here represents the bare minimal reporting that is required to summarise the output for a genetic variation analysis. Whilst we used a trio of individuals to

Figure 1.15: Example iReport: Genetic Report. Example iReport for Clinical Genetics. A) Cover page with custom image. B) First tab, explaining the protocol used. C) Second tab, tables of candidate pathogenic variants, gene columns linking out to GeneCards. D) Fourth tab showing Circos images and family structure.

demonstrate how to select protein-affecting candidate variants based on a recessive model, any number of model outcomes and other assay results may be included in an iReport.

We developed iReport to simplify reporting and sharing the output from *omics* and non-high throughput assays analysed both in and external to Galaxy. We have also utilised iReport for more complex analysis workflows, such as summarising translational research and diagnostic applications for cancer and immunological research and diagnostics.

Availability and requirements

Project name: iReport
Project home page: `https://github.com/shiltemann/iReport`
CTMM-TraIT public Galaxy instance: `https://galaxy.ctmm-trait.nl`
iReport tool shed repository: `https://toolshed.g2.bx.psu.edu/view/saskia-hiltemann/ireport`

Operating system(s): Unix-based Operating Systems
Programming languages: Bash, Perl, Python

Other Requirements: Galaxy
License: GNU GPL
Any restrictions to use by non-academics: none

Examples:
iReport about iReport published history: `http://galaxy.ctmm-trait.nl/u/saskia-hiltemann/h/gcc2014-ireport-about-ireport`, or `tinyurl.com/llrzz9w`
Clinical Genetics iReport published history: `http://galaxy.ctmm-trait.nl/u/andrew-stubbs/h/ireportgeneticreportchr21`

Availability and supporting data

The iReport tool, user manual (published page), and example data and histories are available at the CTMM-TraIT Galaxy server [39].

Abbreviations

CGtag: Complete Genomics Toolkit and Annotation in a Cloud-based Galaxy.
CTMM-TraIT: Center for Translational Molecular Medicine - Translational IT.
NGS: Next Generation Sequencing.
URL: Uniform Resource Locator.

Acknowledgements

This study was performed within the framework of the Center for Translational Molecular Medicine (CTMM). TraIT project (grant 05T-401).

Bibliography

[0] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome biology*, vol. 11, no. 8, p. R86, 2010.

[1] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor, "Galaxy: A web-based genome analysis tool for experimentalists," 2010.

[2] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko, "Galaxy: a platform for interactive large-scale genome analysis," *Genome Res*, vol. 15, no. 10, pp. 1451–1455, 2005.

[3] J. Goecks, C. Eberhard, T. Too, A. Nekrutenko, and J. Taylor, "Web-based visual analysis for high-throughput genomics," *BMC genomics*, vol. 14, no. 1, p. 397, 2013.

[4] S. Hiltemann, H. Mei, M. de Hollander, I. Palli, P. van der Spek, G. Jenster, and A. Stubbs, "CGtag: Complete Genomics toolkit and annotation in a cloud-based Galaxy," *GigaScience*, vol. 3, no. 1, p. 1, 2014.

[5] "Ad hoc reporting." http://reporting.inetsoftware.de/public/remote/adhoc.

[6] "Google charts." https://developers.google.com/chart/.

[7] "Apache openoffice - the free and open productivity suite." https://www.openoffice.org/.

[8] "Bioconductor: open source software for bioinformatics." http://www.bioconductor.org.

[9] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, "Circos: An information aesthetic for comparative genomics," *Genome Res*, vol. 19, no. 9, pp. 1639–1645, 2009.

[10] "Circos circular visualisation." http://circos.ca.

[11] "Omniviz." http://www.instem.com/solutions/omniviz.html.

[12] "Partek: Ngs and microarray data analysis software." https://www.partek.com.

[13] "Tibco spotfire: Business intelligence analytics software and data visualization." https://spotfire.tibco.com.

[14] "Idbs e-workbook." http://www.idbs.com/en/platform-products/e-workbook/inforsense-for-e-workbook/.

[15] "Gensight: Enterprise portfolio management solutions." https://www.gensight.com.

[16] "Cartagenia. confidently interpret, report and share genomic variants." https://www.cartagenia.com.

[17] M. K. Sharma, J. Philips, S. Agarwal, W. S. Wiggins, S. Shrivastava, S. B. Koul, M. Bhattacharjee, C. D. Houchins, R. R. Kalakota, B. George, R. R. Meyer, D. H. Spencer, C. M. Lockwood, T. T. Nguyen, E. J. Duncavage, H. Al-Kateb, C. E. Cottrell, S. Godala, R. Lokineni, S. M. Sawant, V. Chatti, S. Surampudi, R. R. Sunkishala, R. Darbha, S. Macharla, J. D. Milbrandt, H. W. Virgin, R. D. Mitra, R. D. Head, S. Kulkarni, A. Bredemeyer, J. D. Pfeifer, K. Seibert, and R. Nagarajan, "Clinical genomicist workstation," *AMIA Jt Summits Transl Sci Proc*, vol. 19, no. 9, pp. 156–157, 2013.

[18] "Knitr: A general-purpose package for dynamic report generation in r.." http://cran.r-project.org/web/packages/knitr/.

[19] F. Leisch, "Sweave: Dynamic generation of statistical reports using literate data analysis," *Proc in Comp Stat*, pp. 575–580, 2002.

[20] "R-markdown." http://shiny.rstudio.com/articles/rmarkdown.html.

[21] "Tangle." http://worrydream.com/Tangle/.

[22] "Active markdown." http://activemarkdown.org/.

[23] "Ipython notebooks." http://ipython.org/notebook.html.

[24] L. Hermida, C. Poussin, M. B. Stadler, S. Gubian, A. Sewer, D. Gaidatzis, H. R. Hotz, F. Martin, V. Belcastro, S. Cano, M. C. Peitsch, and J. Hoeng, "Confero: an integrated contrast data and gene set platform for computational analysis and biological interpretation of omics data," *BMC Genomics*, vol. 14, p. 514, 2013.

[25] G. Cuccuru, M. Orsini, A. Pinna, A. Sbardellati, N. Soranzo, A. Travaglione, P. Uva, G. Zanetti, and G. Fotia, "Orione, a web-based framework for NGS analysis in microbiology," *Bioinformatics*, vol. 30, pp. 1928–1929, March 2014.

[26] "Galaxy-p." https://usegalaxyp.org/.

[27] "Cgtag pipeline with ireport as final step." http://galaxy.ctmm-trait.nl/u/saskia-hiltemann/p/cgtag.

[28] "ireport example: Tutorial gcc2014." http://galaxy.ctmm-trait.nl/u/saskia-hiltemann/h/gcc2014-ireport-about-ireport(tinyurl.com/llrzz9w).

[29] "ireport example: Genetic report." http://galaxy-demo.trait-ctmm.cloudlet.sara.nl/u/andrew-stubbs/h/ireportgeneticreportchr21.

[30] "jquery zoom library." http://www.jacklmoore.com/zoom/.

[31] "Datatables | table plug-in for jquery." https://datatables.net.

[32] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, "GeneCards: A novel functional genomics compendium with automated data mining and query reformulation support," *Bioinformatics*, vol. 14, no. 8, pp. 656–664, 1998.

[33] "Genecards - the human gene compendium." http://www.genecards.org.

[34] B. MacLean, D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney, B. Frewen, R. Kern, D. L. Tabb, D. C. Liebler, and M. J. MacCoss, "Skyline: an open source document editor for creating and analyzing targeted proteomics experiments," *Bioinformatics*, vol. 26, no. 7, pp. 966–968, 2010.

[35] R. Drmanac, A. Sparks, M. Callow, A. Halpern, N. Burns, B. Kermani, P. Carnevali, I. Nazarenko, G. GB Nilsen, G. Yeung, F. Dahl, A. Fernandez, B. Staker, K. Pant, J. Baccash, A. Borcherding, A. Brownley, R. Cedeno, L. Chen, D. Chernikoff, A. Cheung, R. Chirita, B. Curson, J. Ebert, C. Hacker, R. Hartlage, B. Hauser, S. Huang, Y. Jiang, V. Karpinchyk, and et al., "Human genome sequencing using unchained base reads on self-assembling dna nanoarrays," *Science*, vol. 327, pp. 78–81, January 2010.

[36] The International HapMap Consortium, "The international hapmap consortium. the international hapmap project.," *Nature*, vol. 426, pp. 789–796, 2003.

[37] The International HapMap Consortium, "Integrating common and rare genetic variation in diverse human populations," *Nature*, vol. 467, pp. 52–58, 2010.

[38] "Complete genomics public datasets." http://www.completegenomics.com/public-data/.

[39] "Ctmm-trait public galaxy instance." http://galaxy.ctmm-trait.nl.

"We ignore public understanding of science at our peril"

Eugenie Clark

2

Training

Training is a vital component of accessible research. Galaxy enables the domain experts to perform complex analyses without needing to consult with a bioinformatician. This user-friendliness also makes Galaxy an ideal platform for training, allowing learners to focus solely on the analysis and scientific concepts, rather than the minutiae of tool installation and maintenance or the command line interface details. To this end, we developed a central, community-driven infrastructure for Galaxy training materials, designed for ease-of-use for both trainees and instructors. Furthermore,

by centralising training materials, maintenance and expansion becomes a collaborative effort supported by the global Galaxy trainer community.

This chapter contains the following sub-chapters:

2.0 **Community-Driven Data Analysis Training for Biology**. Together with Bérénice Batut (and later also joined by Helena Rasche), I founded the Galaxy Training Materials (https://training.galaxyproject.org) project in 2016. This involved creating the entire technical framework from the ground up. The web framework utilises Jekyll templating for automatically generating the website from Markdown documents (for both slides and hands-on manuals). This allows for the creation of modern webpages for the tutorials, while allowing tutorial authors to create content in simple Markdown documents. Furthermore, we created the infrastructure needed for the tutorials to be self-contained (e.g. everything needed to follow them is freely and openly available online) and FAIR (e.g. datasets available in Zenodo, automatic BioSchemas annotation for findability in e.g. the TeSS training portal). Through years of active community building efforts, we have grown this into a mature FAIR training platform, currently featuring nearly 200 tutorials across 21 topics (16 scientific and 5 technical), authored by over 170 contributors (https://training.galaxyproject.org/stats). The tutorials typically recreate a published analysis from a scientific journal, further increasing accessibility of the bioinformatics pipelines. These tutorials are also being widely used by teachers around the world, both for bioinformatics workshops for scientists, as well as in higher education curricula. Notable this has also proven to be an invaluable training resource during the current COVID-19 pandemic; due to our focus on FAIR-ness of the materials, our tutorials could be easily used for virtual training events as well as in-person training, exemplified by the recent *GTN Smörgåsbord* training event I organized with Helena Rasche for almost 1200 participants from 78 countries (https://gallantries.github.io/posts/2021/03/01/sm%C3%B6rg%C3%A5sbord/). This chapter is without a doubt the work I am most proud of in this thesis, and continues to be so as we are constantly working to enhance the GTN training platform and grow the community around it.

Community-driven data analysis training for biology

Bérénice Batut[1*], Saskia Hiltemann[2*], Andrea Bagnacani[3], Dannon Baker[7], Clemens Blank[1], Anthony Bretaudeau[4], Loraine Brillet-Guéguen[5], Martin Čech[6], John Chilton[6], Dave Clements[7], Olivia Doppelt-Azeroual[8], Anika Erxleben[1], Mallory Ann Freeberg[9], Simon Gladman[10], Youri Hoogstrate[1], Hans-Rudolf Hotz[11], Torsten Houwaart[1], Pratik Jagtap[12], Delphine Larivière[6], Gildas Le Corguillé[5], Thomas Manke[13], Fabien Mareuil[8], Fidel Ramírez[13], Devon Ryan[13], Florian Christoph Sigloch[1], Nicola Soranzo[14], Joachim Wolff[1], Pavankumar Videm[1], Markus Wolfien[3], Aisanjiang Wubuli[15], Dilmurat Yusuf[1], Galaxy Training Network[16], Rolf Backofen[1], Anton Nekrutenko[6], Björn Grüning[1]

* Bérénice Batut and Saskia Hiltemann contributed equally to this work.

Published in: *Cell Systems*, 2018 Jun 27;6(6):752-758.e1
DOI: https://doi.org/10.1016/j.cels.2018.05.012

1. Albert-Ludwigs-University, Freiburg Germany.
2. Erasmus Medical Centre, Rotterdam, The Netherlands.
3. Department of Systems Biology and Bioinformatics, University of Rostock, Rostock, Germany.
4. INRA, UMR IGEPP, BIPAA/GenOuest, Campus Beaulieu, Rennes, France.
5. CNRS, UMPC, FR2424, ABiMS, Station Biologique, Roscoff, France.
6. The Pennsylvania State University, University Park, PA, USA.
7. Johns Hopkins University, Baltimore MD USA.
8. Bioinformatics and Biostatistics HUB, Centre de Bioinformatique, Biostatistique et Biologie Integrative (C3BI, USR 3756 Institut Pasteur et CNRS) – Paris, France.
9. European Bioinformatics Institute, Hinxton, Cambridge, UK.
10. Melbourne Bioinformatics, The University of Melbourne, Australia.
11. Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland.
12. Biochemistry, Molecular Biology and Biophysics, University of Minnesota Medical School, Minneapolis, USA.

13. Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany.
14. Earlham Institute, Norwich, UK.
15. Leibniz Institute for Farm Animal Biology (FBN), Dummerstorf, Germany.
16. https://galaxyproject.org/teach/gtn/

Abstract

The primary problem with the explosion of biomedical datasets is not the data itself, not computational resources, and not the required storage space, but the general lack of trained and skilled researchers to manipulate and analyze these data. Eliminating this problem requires development of comprehensive educational resources. Here we present a community-developed and driven training framework that enables modern, interactive learning of life sciences data analysis as well as facilitating easy development of tutorials. The key feature of our system is that it is not a static but a continuously improved collection of tutorials. By coupling tutorials with a web-based analysis framework, biomedical researchers can learn by performing computation themselves through a web-browser without the need to install software or search for example datasets. Our ultimate goal is to expand the breadth of training materials to include fundamental statistical and data science topics and to precipitate a complete re-engineering of undergraduate and graduate curricula in life sciences.

Introduction

Rapid development of DNA sequencing technologies has made it possible for biomedical disciplines to rival the physical sciences in data production capability. The combined output of today's sequencing instruments has already surpassed the data generation speed of resources such as the Large Hadron Collider and is rivaling those in the field of astronomy. Yet biology is different from physics (and other quantitative disciplines) in one fundamental aspect—the lack of computational and data analysis training in standard biomedical curricula. Many biomedical scientists do not possess the skills to use or even access existing analysis resources. Such paucity of training also negatively impacts the ability of biomedical researchers to collaborate with their statistics and math counterparts, because of the inability to speak each other's language. In addition, an estimated one-third of biomedical researchers do not have access to proper data analysis support [0]. The only operative way to address these deficiencies is with training. The need for such training cannot be overstated: while the majority (>95%) of researchers work or plan to work with large datasets, most (>65%) possess only minimal bioinformatics skills and are not comfortable with statistical analyses [0], [1]. This overwhelming need drives the demand, which,

at present, greatly exceeds supply [2]. In a recent survey [3] over 60% of biologists expressed a need for more training while only 5% called for more computing power. Thus one can assume that the true bottleneck of the current data deluge is not storage or processing power, but the knowledge and skills to utilize existing resources.

Since 2006 our team has been pondering the question of how to enable computationally novice users to perform complex data analysis tasks. We attempted to solve this problem by creating a platform, Galaxy (`http://galaxyproject.org` [4]), that provides access to hundreds of tools used in a wide variety of analysis scenarios. It features a web-based user interface while automatically and transparently managing underlying computation details [4]. It can be deployed on a personal computer, heterogeneous computer clusters, as well as computation systems provided by Amazon, Microsoft, Google and other clouds such as those running OpenStack. Over the years a community has formed around this project, providing it with an ever-growing, up-to-date set of analysis tools and expanding it beyond life sciences.

These features of Galaxy attracted many biomedical researchers, making it well suited for use as a teaching platform. Here we describe a community-driven effort to build, maintain, and promote a training infrastructure designed to provide computational data analysis training to biomedical researchers worldwide. Our effort utilizes the Galaxy platform [4] to support a comprehensive training portfolio and relies on modern web-based technologies for content maintenance and delivery.

Results and Discussion

Our goal is to develop an infrastructure that facilitates data analysis training in life sciences. At a minimum it needs to provide an interactive learning platform tuned for current datasets and research problems. It should also provide means for community-wide content creation and maintenance, and, finally, enable trainers and trainees to use the tutorials in a variety of situations such as those where reliable Internet access is not an option.

Interactive learning tailored to research problems

Our main result is a collection of hands-on tutorials that are designed to be interactive and are built around Galaxy. The hands-on nature of our training material implies that a trainee can have two web browser windows open side-by-side: one pointed at the current tutorial and the other at a Galaxy instance. We build most tutorials around a "research story": a scenario inspired by a

Topic	Target	Tutorials
Galaxy Server administration	Admin	Galaxy Database schema, Docker and Galaxy, Advanced customisation of a Galaxy instance
Assembly	Biol	Introduction to Genome Assembly, De Bruijn Graph Assembly, Unicycler Assembly
ChIP-Seq data analysis	Biol	Identification of the binding sites of the T-cell acute lymphocytic leukemia protein 1 (TAL1), Identification of the binding sites of the Estrogen receptor
Development in Galaxy	Dev	Contributing with GitHub, Tool development and integration into Galaxy, Tool Shed: sharing Galaxy tools, Galaxy Interactive Tours, Galaxy Interactive Environments, Visualizations: charts plugins, Galaxy Webhooks, Visualizations: generic plugins, BioBlend module, a Python library to use Galaxy API, Tool Dependencies and Conda, Tool Dependencies and Containers, Galaxy Code Architecture
Epigenetics	Biol	DNA Methylation
Introduction to Galaxy	Biol	Galaxy 101, From peaks to genes, Multisample Analysis, Options for using Galaxy, IGV Introduction, Getting data into Galaxy
Metagenomics	Biol	16S Microbial Analysis with Mothur, Analyses of metagenomics data - The global picture
Proteomics	Biol	Protein FASTA Database Handling, Metaproteomics tutorial, Label-free versus Labelled - How to Choose Your Quantitation Method, Detection and quantitation of N-termini via N-TAILS, Peptide and Protein ID, Secretome Prediction, Peptide and Protein Quantification via Stable Isotope Labelling (SIL)
Sequence Analysis	Biol	Quality Control, Mapping, Genome Annotation, RAD-Seq Reference-based data analysis, RAD-Seq de-novo data analysis, RAD-Seq to construct genetic maps
Train the trainers	Inst	Creating a new tutorial - Writing content in markdown, Creating a new tutorial - Defining metadata, Creating a new tutorial - Setting up the infrastructure, Creating a new tutorial - Creating Interactive Galaxy Tours, Creating a new tutorial - Building a Docker flavor for a tutorial, Good practices to run a workshop
Transcriptomics	Biol	De novo transcriptome reconstruction with RNA-seq, Reference-based RNA-seq data analysis, Differential abundance testing of small RNAs

Table 2.0: Topics available in the Galaxy training material website (https://training.galaxyproject.org) with their target users and available tutorials. Admin = Galaxy administrators, Biol = Biomedical researchers, Dev = Tool and software developers, Inst = Instructors and Tutorial developers. The scripts for the extraction of such information are available in GitHub (https://github.com/bebatut/galaxy-training-material-stats). This table displays content current as of 28th Sep, 2017.

previously published manuscript or an interesting dataset (with the caveat that some more technical materials do not lend themselves to this goal). To make training comprehensive, we aim to cover major branches of biomedical big data applications such as those listed in Table 2.0.

As an example, suppose that a researcher is interested in learning about metagenomic data analyses. The category "Metagenomics" at `https://training.galaxyproject.org` presently contains a set of introductory slides, two hands-on tutorials (Fig. 2.0A), and HTML-based slides designed as a brief (10 - 20 min) introduction to the subject. In addition, every hands-on tutorial begins with background information (Fig. 2.0B) and explains how it influences data analysis. This background story is included to account for situations when tutorials are used for self-teaching in the absence of an instructor who would provide a formal introduction. After the introduction, the hands-on part of the tutorial begins and is laid out in a step-by-step fashion with explanations (boxes on Figure 2.0D) of what is being done inside Galaxy, which parameters are critical, and how modifying parameters affects downstream results. The first step in this progression is usually a description of the datasets and how to obtain them. We invested a large effort in creating appropriate datasets by downsampling original published data, which is necessary since real-world datasets are usually too big for tutorials. Our goal was to make datasets as small as possible while still producing an interpretable result. We use Zenodo (`http://www.zenodo.org`), an open data archiving and distribution platform, to store the tutorial datasets and to provide them with stable digital object identifiers (DOIs) that can be used to credit their authors and for citation purposes.

Tutorials start with a list of prerequisites (typically other tutorials within the site) to account for the variation in trainees' backgrounds, a rough time estimate, questions addressed during the tutorial, learning objectives, and key points (e.g., Fig. 2.0D). These components help trainees and instructors to keep track of the training goals. For example, the learning objectives are single sentences describing what a trainee will be able to do as a result of the training [5]. Throughout the tutorials, question boxes (Fig. 2.0C) are added as an effective way to motivate the trainees [6] [7] and guide self-training. The training material is distributed under a CC BY 4.0 (`https://creativecommons.org/licenses/by/4.0/`) license: its contents can be shared and adapted as long as appropriate credit is given. Efforts have been made also in the direction of ensuring website accessibility to disabled persons by regular evaluation with WAVE (`http://wave.webaim.org`), a web accessibility evaluation tool, and by automatic checking for alternative text for the images.

Keeping trainees engaged is critical, particularly for self-training. To this end, we aim to provide interactive tours for each tutorial: using instruction bubbles (Fig. 2.0E), each tutorial step can be

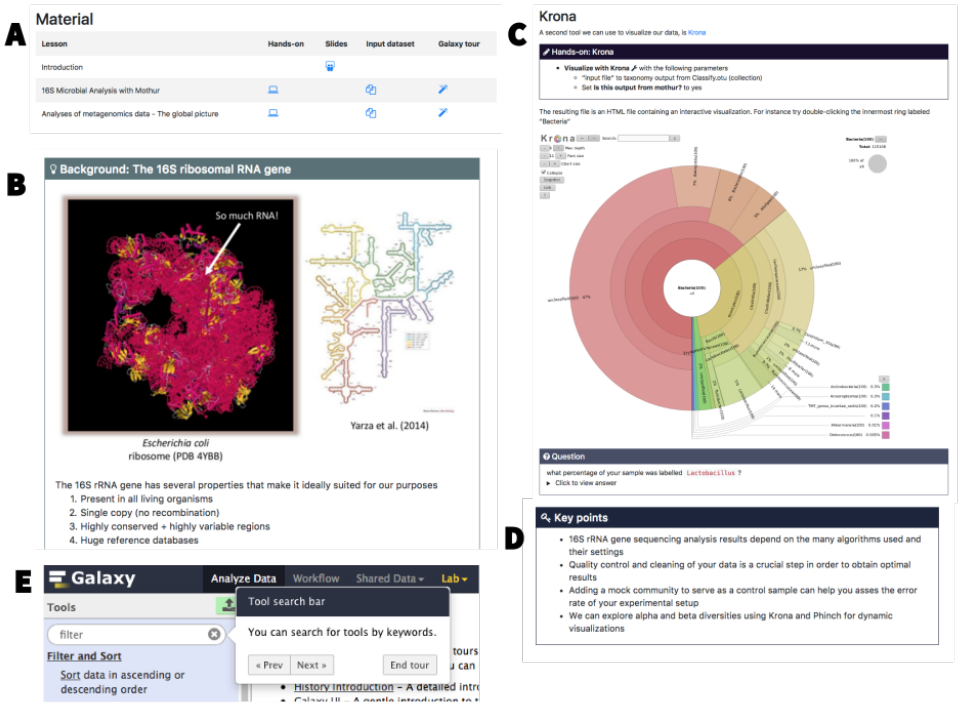


Figure 2.0: Key elements of an interactive tutorial. **A.** A list of tutorials dedicated to Metagenomics. There is a set of introductory slides and two hands-on tutorials. **B.** A fragment of introductory material within a tutorial. **C.** A "hands-on" element with upper box contains instructions for running a tool inside Galaxy and shown example output of Krona tool. The question box at the bottom contains togglable answer field. **D.** Summary of key points for this tutorial displayed on the bottom of the tutorial. **E.** Fragment of Galaxy interface showing interactive tour balloon.

"played" directly inside Galaxy, guiding learners to the needed tools while also allowing exploration of the framework's functionalities.

Infrastructure to facilitate community-led content development

To build a comprehensive collection of training materials covering the spectrum of topics in the life sciences, we must leverage community expertise, as no single group can possibly "know it all". To achieve this goal, we built an infrastructure that makes tutorial creation a convenient, hassle-free process and enables transparent peer-review and curation to guarantee high-quality and current content. In implementing these requirements, we took inspiration from the Software and Data Carpentry [8] projects (SDC). In SDC, materials are openly reviewed and iteratively developed on GitHub (`https://github.com/`) to capture the breadth of community expertise. SDC delivers training via online tutorials with hands-on sections, which offer better training support

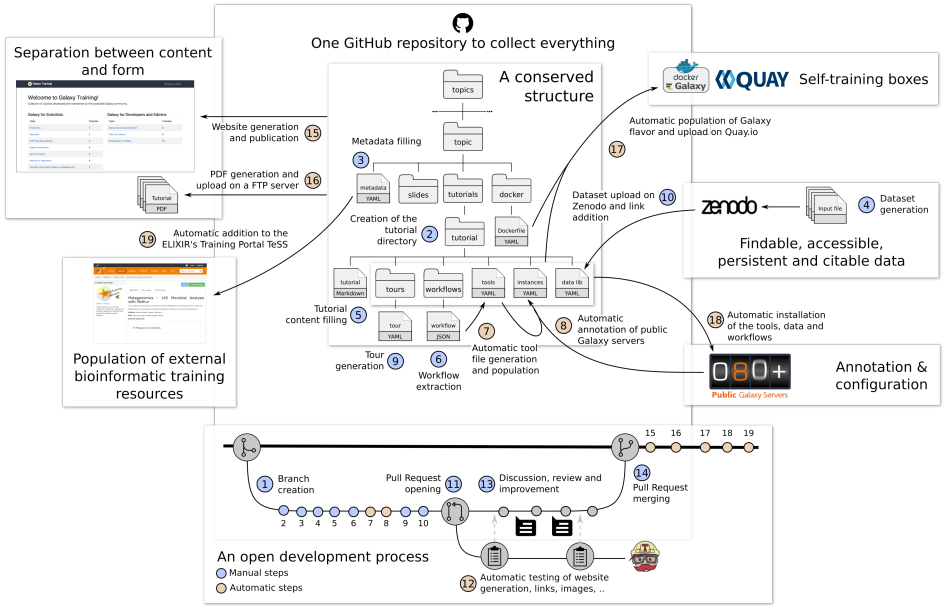


Figure 2.1: Structure and development of content in GitHub (http://github.com/galaxyproject/training-material). The material is organized in different topics, each topic in a dedicated directory. Inside each topic's directory, the structure is the same: a metadata file, a directory with the topic introduction slide decks, a directory with the tutorials and a directory with the Dockerfile describing the details to build a container for the topic that would contain a dedicated Galaxy instance with all tools relevant for the tutorials. Inside the topic directory, each tutorial related to the topic has its own subdirectory with several files: a tutorial file written in Markdown with hands-on, an optional slides file to support the tutorial, a directory with Galaxy Interactive Tours to reproduce the tutorial, a directory with workflows extracted from the tutorial, a file with the links to the input data needed for the tutorial and a file with the description of needed tools to run the tutorial. The process of development of new content is shown at the bottom of the figure.

than videos because trainees who are actively participating learn more [6]. This format is also adapted to face-to-face courses and self-training, as the content is openly accessible online. The content of these web pages is easy to edit, thus reducing the contribution barrier. The tutorials are developed in Markdown, a plain text markup language, which is automatically transformed into web-browser accessible pages. Using these strategies, we created a GitHub repository (`https://github.com/galaxyproject/training-material`) to collect, manage, and distribute training materials. The architecture of this infrastructure is shown in Fig. 2.1 (center), with the process for developing a tutorial illustrated at the bottom of the figure. To create a new tutorial, the main repository is forked (duplicated into a user-controlled space) within GitHub by an individual developing the tutorial. The developer then proceeds to write the content using Markdown as explained in our guide at `https://training.galaxyproject.org/topics/training` (itself

consisting of several tutorials). The guide contains detailed information on technical and stylistic aspects of tutorial development. After settling on a final version of the tutorial (circles 1 through 10, the bottom of Fig. 2.1), a pull request is created against the original repository. When a new pull request is issued, this is an indication that a new tutorial is ready to be reviewed by the editorial team. The team then makes suggestions on the new contents, these suggestions are discussed, and the content is edited accordingly. A decision is then made whether to accept the pull request. At the same time the pull request is first created, the newly added content is automatically tested for HTML generation and all links and images are verified. When the pull request is accepted, the new tutorial becomes a part of the official training material portfolio, and the entire site is regenerated.

This infrastructure has been developed in accordance with the FAIR (Findable, Accessible, Interoperable, Reusable) principles [9]. Each tutorial, slide deck, and topic is complemented by numerous metadata described in a standard, accessible, interoperable format (YAML; `http://yaml.org/`). The metadata is used to automatically populate the TeSS training portal at the European life-sciences Infrastructure for biological Information (ELIXIR; `https://tess.elixir-europe.org`), ensuring global reach [10]. Each topic, tutorial, and slide deck has as metadata a reference to a topic in the EDAM ontology [11], a comprehensive catalog of well-established, familiar concepts that are prevalent within bioinformatics and computational biology. These references can be used to represent relationships among the materials and make them more findable and searchable.

Using the framework described above, we relaunched the Galaxy Training Network (GTN; `https://galaxyproject.org/teach/gtn`). This growing network currently consists of 33 scientific groups (`https://galaxyproject.org/teach/trainers`) invested in Galaxy-based training. The GTN regularly organizes training events worldwide (Fig. 2.2) and offers best practices for developing Galaxy-based training material, advice on compute platform choice to use for training, and a catalog of existing training resources for Galaxy (Table 2.0).

Ensuring accessibility of tutorials

Most training materials hosted within the GTN resource are intended to be used side-by-side with the Galaxy framework. However, the public Galaxy instances (e.g. `https://usegalaxy.org` or `https://galaxy.uni-freiburg.de`) are occasionally subject to unpredictable load, may be inaccessible due to network problems in remote parts of the world or may not have all the tools necessary for completing the tutorials. To account for these situations, we have developed a Docker-based framework for creating portable, on-demand Galaxy instances specifically targeted for a given tutorial. Docker (`https://www.docker.com`) is a container platform which provides

lightweight virtualization by executing "images" (files that include everything needed to run a piece of software) isolated from the host computer environment. An individual creating a new tutorial lists all tools that are required to complete it in a dedicated configuration file (tools file, Fig. 2.1). For example, a metagenomics tutorial uses the mothur [12] set of tools as well as visualization applications such as Krona [13]. The corresponding Galaxy tools are listed in a configuration file that is a part of the metagenomics tutorial. This file is used to install these Galaxy tools and their dependencies into a base Galaxy Docker image (containing essential Galaxy functionality and a core set of tools) to create a dedicated "on-demand" Galaxy instance which can then be used on any trainer's or trainee's computer. The Docker image also contains input data, tours, workflows.

A vision for the future

Life sciences are on a trajectory towards becoming an entirely data-driven scientific domain. A growing understanding that biomedical curricula must be modernized to reflect these changes is gaining attention [14]. Our project represents one of the first fully open, "grass-roots" attempts at unifying and standardizing heterogeneous training resources around the Galaxy platform. While it may not be appropriate to all, our multi-year experience with teaching workshops at various skill-levels can be summarized as the following set of recommendations, which we use as guiding principles. These recommendations may also be useful for the development of alternative frameworks as well as for curriculum planning:

1. **Require quantitative training.** No one expects biomedical researchers to rival their colleagues in departments of mathematics or statistics. However, background level statistical reasoning must be included in all training materials and general statistical courses must become a part of undergraduate and graduate education. This would have an enormous positive impact on the quality of biomedical research because researchers with basic understanding of quantitative concepts will not, for example, perform an RNAseq experiment without a sufficient number of replicates.

2. **Demystify computational methodologies.** Fundamental principles, limitations, and assumptions of molecular experimental techniques are typically well understood by biomedical researchers even when proprietary reagent kits are used. This is not the case with software tools, which are often treated as black boxes. We argue that fundamental principles of bioinformatic techniques (e.g., read mapping, read assembly) must be understood by experimentalists as this will also lead to an increase in overall quality of research output.

3. **Advocate the fundamental virtues of open and transparent research.** Open and transparent data analysis (e.g., through the use of open-source software) promotes replication and validation of results by independent investigators. It also speeds up research progress by facilitating reuse and repurposing of published analyses to different datasets or even to other disciplines. We advocate openness as a basic principle for computational analysis of biomedical data.

The infrastructure presented here has been developed to support training using Galaxy, a powerful tool for teaching bioinformatics concepts and analysis. But such a model is not only limited to Galaxy. It could be applied to bioinformatics training more generally (and to other disciplines as well) to support learners and instructors in this ever-changing landscape that is the life sciences.

Acknowledgments

The authors are grateful to the Freiburg Galaxy and Core Galaxy teams, as without these resources this work would not be possible. Adoption of Galaxy Tours has been accelerated with the introduction of Galaxy Tour Builder (https://zenodo.org/record/830481) by William Durand (https://tailordev.fr). This project was supported by Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012), German Federal Ministry of Education and Research (BMBF grant 031 A538A RBC (de.NBI)), NIH Grants U41 HG006620 and R01 AI134384-01, as well as NSF Grant 1661497. References

Bibliography

[0] L. Larcombe, R. Hendricusdottir, T. Attwood, F. Bacall, N. Beard, L. Bellis, W. Dunn, J. Hancock, A. Nenadic, C. Orengo, *et al.*, "Elixir-uk role in bioinformatics training at the national level and across elixir," *F1000Research*, vol. 6, 2017.

[1] J. J. Williams and T. K. Teal, "A vision for collaborative training infrastructure for bioinformatics," *Annals of the New York Academy of Sciences*, vol. 1387, no. 1, pp. 54–60, 2017.

[2] T. K. Attwood, S. Blackford, M. D. Brazas, A. Davies, and M. V. Schneider, "A global perspective on evolving bioinformatics and data science training needs," *Briefings in Bioinformatics*, vol. 20, pp. 398–404, Aug. 2017.

[3] "Community survey report - 2013 - embl-abr." https://www.embl-abr.org.au/news/braembl-community-survey-report-2013/.

[4] E. Afgan, D. Baker, M. Van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, *et al.*, "The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W3–W10, 2016.

[5] A. Via, T. Blicher, E. Bongcam-Rudloff, M. D. Brazas, C. Brooksbank, A. Budd, J. De Las Rivas, J. Dreyer, P. L. Fernandes, C. Van Gelder, *et al.*, "Best practices in bioinformatics training for life scientists," *Briefings in bioinformatics*, vol. 14, no. 5, pp. 528–537, 2013.

[6] A. Dollár, P. S. Steif, and R. Strader, "Enhancing traditional classroom instruction with web-based statics course," in *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual*, pp. F1H–1, IEEE, 2007.

[7] R. Scheines, G. Leinhardt, J. Smith, and K. Cho, "Replacing lecture with web-based course materials," *Journal of Educational Computing Research*, vol. 32, no. 1, pp. 1–25, 2005.

[8] G. Wilson, "Software carpentry: lessons learned," *F1000Research*, vol. 3, 2014.

[9] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, p. 160018, 2016.

[10] N. Beard, T. Attwood, and A. Nenadic, "Tess – training portal," July 2016.

[11] J. Ison, M. Kalaš, I. Jonassen, D. Bolser, M. Uludag, H. McWilliam, J. Malone, R. Lopez, S. Pettifer, and P. Rice, "Edam: an ontology of bioinformatics operations, types of data and identifiers, topics and formats," *Bioinformatics*, vol. 29, no. 10, pp. 1325–1332, 2013.

[12] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, *et al.*, "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and environmental microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[13] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Krona: interactive metagenomic visualization in a web browser," *Encyclopedia of Metagenomics: Genes, Genomes and Metagenomes: Basics, Methods, Databases and Tools*, pp. 339–346, 2015.

[14] P. Hitchcock, A. Mathur, J. Bennett, P. Cameron, C. Chow, P. Clifford, R. Duvoisin, A. Feig, K. Finneran, D. M. Klotz, R. McGee, M. O'Riordan, C. Pfund, C. Pickett, N. Schwartz, N. E. Street, E. Watkins, J. Wiest, and D. Engelke, "Point of view: The future of graduate and postdoctoral training in the biosciences," *eLife*, vol. 6, p. e32715, oct 2017.

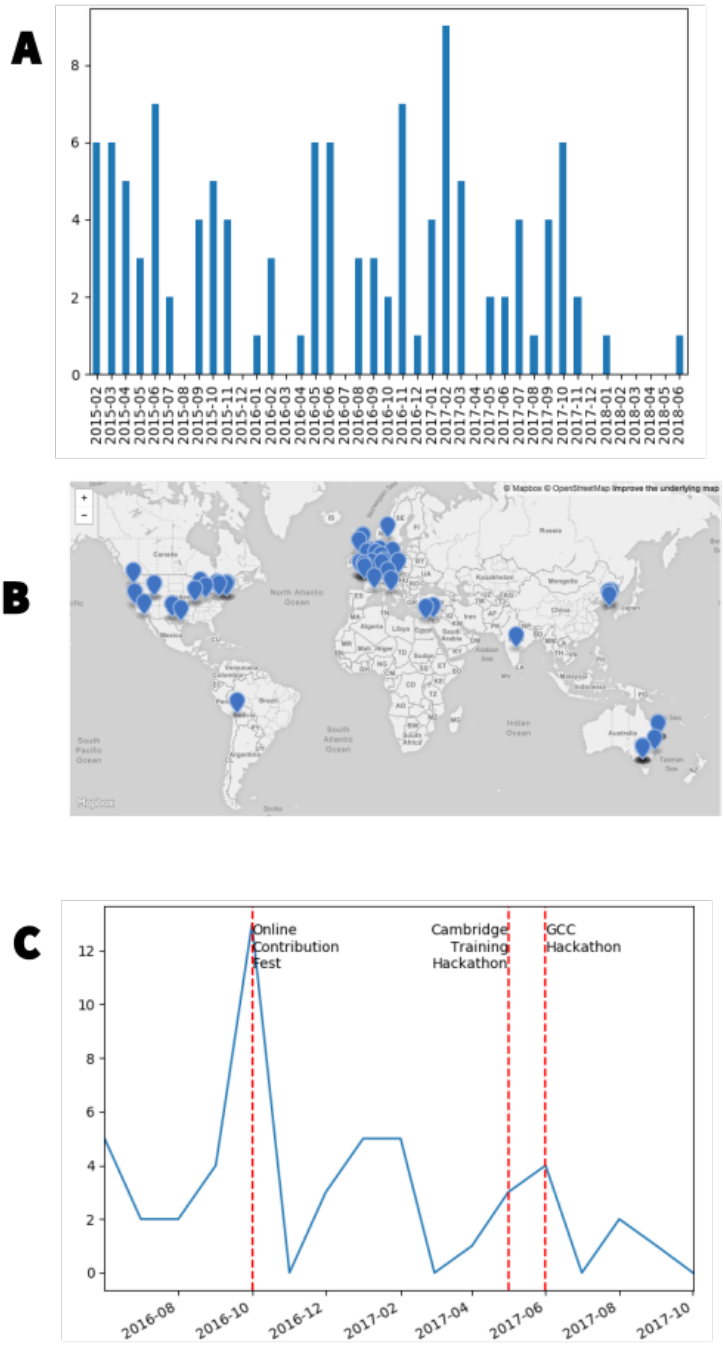
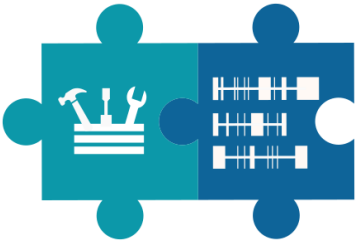


Figure 2.2: History of training activities. Number (A) and location (B) of registered training events organized by the Galaxy Training Network since 2015. C. Number of tutorial contributors per month.



"As always in life, people want a simple answer and it's always wrong."

Susan Greenfield

3

Structural Variant Analysis

Structural variations are large-scale rearrangements of the genome. When these alterations occur within genes, novel hybrid genes called *fusion genes* may be formed. Accurate detection of SVs and resulting fusion genes are an informative in cancer research studies.

We created iFUSE, web-based application to visualize and explore structural variants, and identify and prioritize potential fusion genes within a sample. We subsequently used this tool

to detect multiple fusions in VCaP cell line. Futhermore, we were discovered chromothirpsis on chromosome 5q of this sample, and used the Circos tool to visualize this in a single plot.

This chapter contains the following sub-chapters:

3.0 **iFUSE: integrated fusion gene explorer** In this work, we created an interactive web-based visualisation tool for the identification and prioritization of fusion gene candidates. Coding of the application was done by Jos van Nijnatten (PHP web framework), Elizabeth McClellan (R data analysis), and later finalized, adapted by myself (PHP, R & bash). I also worked to increase the interoperability of the code, extending it to additional file formats beyond Complete Genomics. Together with Ines Teles Alves, I worked to validate the application. Ivo Palli made sure the application was available as a robust web service to visitors from outside the EMC. However, due to changed policies at the EMC in the intervening years, we are no longer allowed to offer the iFUSE application as a service to users from outside the EMC network. Therefore the link mentioned in this work is no longer functional. In order to keep iFUSE freely available to all, I have created a Docker image of the application, allowing anybody to run iFUSE on their own computer. Please see the GitHub repository at `https://github.com/erasmusmc-bioinformatics/ifuse` for further details.

3.1 **Gene fusions by chromothripsis of chromosome 5q in the VCaP prostate cancer cell line** In this work, we used the iFUSE application to identify and validate fusion gene candidates. I led the bioinformatics analysis while Ines Teles Alves led the experimental validation. While we analyzed many samples, we came across a surprise in the VCaP sample, it seemed to display chromothripsis on the q arm of chromosome 5. We were able to confirm this through a combination of bioinformatic analysis and experimental validation.

Note on Availability

iFUSE: integrated fusion gene explorer

Saskia Hiltemann [1], Elizabeth A. McClellan[2], Jos van Nijnatten[2], Sebastiaan Horsman[2], Ivo Palli[2], Ines Teles Alves[1,3], Thomas Hartjes[1], Jan Trapman[3], Peter van der Spek[2], Guido Jenster[1], and Andrew Stubbs[2]

1. Department of Urology, Erasmus MC, 3015 GE Rotterdam, The Netherlands.
2. Department of Bioinformatics, Erasmus MC, 3015 GE Rotterdam, The Netherlands.
3. Department of Pathology, Erasmus MC, 3015 GE Rotterdam, The Netherlands.

Published in: *Bioinformatics*, 2013 Jul 1;29(13):1700-1
DOI: https://doi.org/10.1093/bioinformatics/btt252

Abstract

Summary: We present iFUSE (integrated FUSion gene Explorer), an online visualization tool that provides a fast and informative view of structural variation data and prioritizes those breaks likely representing fusion genes. This application uses calculated breakpoints to determine fusion genes based on the latest annotation for genomic sequence information, and where relevant the structural variation (SV) events are annotated with predicted RNA and protein sequences. iFUSE takes as input either a Complete Genomics (CG) junction file, a FusionMap [o] fusion detection report file, or a file already analysed and annotated by the iFUSE application on a previous occasion.

Results: We demonstrate the utility of iFUSE with case studies from tumour-normal SV detection derived from Complete Genomics whole-genome sequencing results.

Availability: iFUSE is available as a web service at http://ifuse.erasmusmc.nl

Introduction

Structural variation analysis is a common requirement in the study of cancer where many fusion genes have been implicated in the progression of cancer [1, 2] The use of next-generation sequencing for fusion gene detection in cancer [3, 0, 4], structural variation in non-cancerous diseases [5, 6] and in normal genomes [7] has expanded knowledge of the importance of these events. In a recent study the use of *de novo* assembly in association with SV detection suggests that SVs account for more diversity between individuals than do single nucleotide polymorphisms (SNPs) [8]. Complete Genomics uses *de novo* assembly during SV, single nucleotide variation (SNV) and copy number variation (CNV) determination [9].

Traditionally, for a given SV, a user can visualize the individual (sides of the) breaks in viewers such as IGV [10, 11], but not the resulting event or fusion gene as a whole, and must manually retrieve the sequence of the resulting event based on the exons from both genes of the proposed fusion gene and determine the orientation and frame of the predicted transcript and encoded polypeptide sequences. Other applications also offer visualisation, but with the caveat that the user must process their data within the application, e.g inGAP-SV [12], or that a bioinformatician is required to render the visualization using e.g. Circos [13]. Our aim is to deliver a web-based application that allows scientists to visualize all detected SV events and fusion genes determined in their results and to provide the concomitant candidate transcripts and polypeptide sequences associated with the detected fusion genes (Figure 3.0). No other application exists at the moment to categorize and visualize the candidate fusion genes, and determine the proposed sequence for gDNA, RNA and polypeptides from standard SV files.

Methods

iFUSE is a PHP-coded application running on an Apache web server with a mySQL database for user management and R for data analysis. Gene-based feature annotation is provided using the UCSC gene tables (hg18 and hg19). Documentation details, including full application configuration, are available from the website (http://ifuse.erasmusmc.nl).

iFUSE takes as input either a Complete Genomics (CG) junctions file or a fusion detection report as generated by the FusionMap tool [0]. To perform a comparison of two or more genomes, the Complete Genomics tool Junctions2Events can be used prior to visualisation within iFUSE. The input file is validated and then analyzed using R, after which a graphical representation for each event is generated. This representation displays the promotor, introns, exons and the junction site, and additional information including DNA, RNA and protein sequences are presented to the user. These events can be filtered and sorted by the user, either using general properties or by zooming in on a single event and filtering for nearby junctions or events with similar properties.

The input files uploaded to iFUSE are annotated in R using information retrieved from UCSC gene tables and the input files. The resulting output file can be retrieved from iFUSE for manual inspection, and can also be

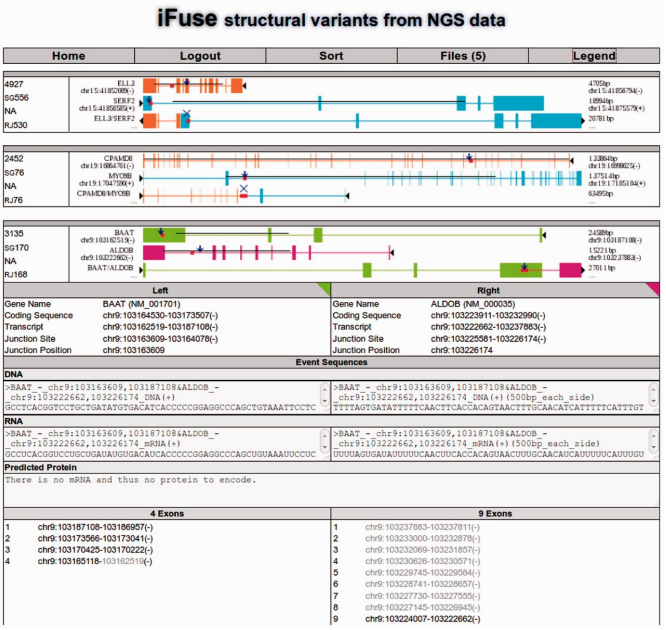


Figure 3.0: Screenshot of iFUSE. SVs are visualised and where applicable, the predicted DNA, RNA and protein sequences are reported.

used directly as input to iFUSE.

Example input files can be downloaded from the downloads section of the iFUSE website and users can test the application without registration by selecting the option to start an anonymous session. Any files uploaded by the user will be deleted at the end of the anonymous session.

iFUSE accounts are password protected, the application has been tested for security risks such as SQL injections, and our servers undergo periodical CERT vulnerability scans (http://www.cert.org). Furthermore, when a user deletes a file via the iFUSE web interface, it is purged completely from our systems.

Discussion

Two public cancer genomes have been used to demonstrate the utility of iFUSE for the prediction of fusion genes. Genomes were downloaded from Complete Genomics (ftp2.completegenomics.com). The results can be found in Table 3.0.

An event is labeled as a fusion gene if the breakpoint has two different genes on either side. If the two sides also have the same orientation (are on the same strand, or in the case of an inversion on opposite strands) and

	S1 Tumour	S1 Normal	S2 Tumour	S2 Normal
HG18				
Junctions	1579	1594	1558	1387
Genes on both sides	32	15	23	4
Same orientation	21	14	16	1
HG19				
Junctions	1581	1592	1559	1390
Genes on both sides	21	12	31	7
Same orientation	10	6	17	3

Table 3.0: Results from iFUSE on public datasets. Sample 1 (S1): HCC1187, Sample 2 (S2): HCC2218, public datasets downloadable from Complete Genomics. (ftp://ftp2.completegenomics.com/Cancer_pairs/)

are also in frame, the event is called a fusion protein.

Conclusion

iFUSE provides scientists with a convenient method to visualize, categorize, and filter structural variation analysis output using Complete Genomics junction files or the FusionMap fusion detection report files as input to the application.

Acknowledgements

This study was performed within the framework of CTMM, the Center for Translational Molecular Medicine. TraIT project (grant 05T-401).

We would like to thank Rick Tearle and Steve Lincoln from Complete Genomics whose valuable discussions on Complete Genomics analysis methods supported our application development.

Bibliography

[0] H. Ge, K. Liu, T. Juan, F. Fang, M. Newman, and W. Hoeck, "Fusionmap: detecting fusion genes from next-generation sequencing data at base-pair resolution," *Bioinformatics*, vol. 27, no. 14, pp. 1922–1928, 2011.

[1] F. Mitelman, B. Johansson, and F. Mertens, "The impact of translocations and gene fusions on cancer causation," *Nature Reviews Cancer*, vol. 7, no. 4, pp. 233–245, 2007.

[2] C. Kumar-Sinha, S. A. Tomlins, and A. M. Chinnaiyan, "Recurrent gene fusions in prostate cancer," *Nature Reviews Cancer*, vol. 8, no. 7, pp. 497–511, 2008.

[3] H. Edgren, A. Murumagi, S. Kangaspeska, D. Nicorici, V. Hongisto, K. Kleivi, I. H. Rye, S. Nyberg, M. Wolf, A.-L. Borresen-Dale, *et al.*, "Identification of fusion genes in breast cancer by paired-end rna-sequencing," *Genome biology*, vol. 12, no. 1, p. R6, 2011.

[4] A. McPherson, F. Hormozdiari, A. Zayed, R. Giuliany, G. Ha, M. G. Sun, M. Griffith, A. H. Moussavi, J. Senz, N. Melnyk, *et al.*, "defuse: an algorithm for gene fusion discovery in tumor rna-seq data," *PLoS computational biology*, vol. 7, no. 5, p. e1001138, 2011.

[5] S. J. Sanders, A. G. Ercan-Sencicek, V. Hus, R. Luo, M. T. Murtha, D. Moreno-De-Luca, S. H. Chu, M. P. Moreau, A. R. Gupta, S. A. Thomson, *et al.*, "Multiple recurrent de novo cnvs, including duplications of the 7q11. 23 williams syndrome region, are strongly associated with autism," *Neuron*, vol. 70, no. 5, pp. 863–885, 2011.

[6] D. Levy, M. Ronemus, B. Yamrom, Y.-h. Lee, A. Leotta, J. Kendall, S. Marks, B. Lakshmi, D. Pai, K. Ye, *et al.*, "Rare de novo and transmitted copy-number variation in autistic spectrum disorders," *Neuron*, vol. 70, no. 5, pp. 886–897, 2011.

[7] . G. P. Consortium *et al.*, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[8] Y. Li, H. Zheng, R. Luo, H. Wu, H. Zhu, R. Li, H. Cao, B. Wu, S. Huang, H. Shao, *et al.*, "Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly," *Nature biotechnology*, vol. 29, no. 8, pp. 723–730, 2011.

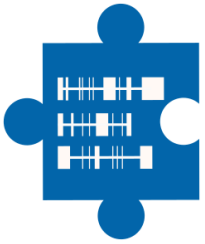
[9] P. Carnevali, J. Baccash, A. L. Halpern, I. Nazarenko, G. B. Nilsen, K. P. Pant, J. C. Ebert, A. Brownley, M. Morenzoni, V. Karpinchyk, *et al.*, "Computational techniques for human genome resequencing using mated gapped reads," *Journal of Computational Biology*, vol. 19, no. 3, pp. 279–292, 2012.

[10] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, "Integrative genomics viewer (igv): high-performance genomics data visualization and exploration," *Briefings in bioinformatics*, vol. 14, no. 2, pp. 178–192, 2013.

[11] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov, "Integrative genomics viewer," *Nature biotechnology*, vol. 29, no. 1, pp. 24–26, 2011.

[12] J. Qi and F. Zhao, "ingap-sv: a novel scheme to identify and visualize structural variation from paired end mapping data," *Nucleic acids research*, vol. 39, no. suppl_2, pp. W567–W575, 2011.

[13] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, "Circos: an information aesthetic for comparative genomics," *Genome research*, vol. 19, no. 9, pp. 1639–1645, 2009.



Gene fusions by chromothripsis of chromo- some 5q in the VCaP prostate cancer cell line

Inês Teles Alves[2,3,*], Saskia Hiltemann[1,2,*], Thomas Hartjes[2], Peter van der Spek[1], Andrew Stubbs[1], Jan Trapman[3], Guido Jenster[2]

1. Department of Bioinformatics, Erasmus Medical Center, Rotterdam, The Netherlands.
2. Department of Urology, Josephine Nefkens Institute, Erasmus Medical Center, Rotterdam, The Netherlands.
3. Department of Pathology, Josephine Nefkens Institute, Erasmus Medical Center, Rotterdam, The Netherlands.

Published in: *Human Genetics*, 2013 Jun;132(6):709-13

DOI: https://doi.org/10.1007/s00439-013-1308-1

* Inês Teles Alves and Saskia Hiltemann contributed equally to this work.

Supplementary material available online.

Abstract

The VCaP cell line is widely used in prostate cancer research as it is a unique model to study castrate resistant disease expressing high levels of the wild type androgen receptor and the *TMPRSS2-ERG* fusion transcript. Using next generation sequencing, we assembled the structural variations in VCaP genomic DNA and observed a massive number of genomic rearrangements along the q arm of chromosome 5, characteristic of chromothripsis. Chromothripsis is a recently recognized phenomenon characterized by extensive chromosomal shattering in a single catastrophic event,

mainly detected in cancer cells. Various structural events identified on chromosome 5q of VCaP resulted in gene fusions. Out of the 18 gene fusion candidates tested, 15 were confirmed on genomic level. In our set of gene fusions, only rarely we observe microhomology flanking the breakpoints. On RNA level, only five transcripts were detected and *NDUFAF2-MAST4* was the only resulting in an in-frame fusion transcript. Our data indicate that although a marker of genomic instability, chromothripsis might lead to only a limited number of functionally relevant fusion genes.

Report

Advances in DNA sequencing technologies have allowed the detailed analysis of genomic aberrations in cancer. Stephens et al. (2011) [0], described massive genomic rearrangements, designated 'chromothripsis', in a subgroup of chronic lymphocytic leukemia. Subsequently, chromothripsis has been observed in several cancer cell lines, including lung, sarcoma, esophageal, renal, and thyroid cells and in a few patient samples of multiple myeloma, colorectal cancer, medulloblastoma, and neuroblastoma [1, 2, 3, 4].

The process of chromothripsis involves the shattering of one or a few chromosomes that is followed by fragment reassembly into derivative chromosomes [5]. It is defined on the basis of three main features: the occurrence of numerous genomic rearrangements in localized chromosomal regions; several copy number changes alternating between only one, two, or occasionally three different copy number states; and the alternation between regions where heterozygosity is preserved with regions displaying loss of heterozygosity [5]. The localized pattern observed for chromothripsis differs from that of other types of genomic instability, where rearrangements tend to be dispersed genome-wide [6, 7] and suggests chromothripsis will likely occur when chromosomes are largely condensed. Moreover, the alternation between few copy number states strongly implies that the chromosomal rearrangements occurred in a short time scale, probably in one single mutational event [0, 8].

Several mechanisms have been proposed to induce the massive number of genomic rearrangements observed in chromothripsis. Overall, the clustering pattern of rearrangements observed in chromothripsis is readily explained by assuming a condensed configuration of the chromosome by the time chromothripsis was triggered. Moreover, one can also consider, although less reasonably, that a chromosomal region is exposed to localized high-energy ionizing radiation [9]. Along with telomere erosion and breakage–fusion–bridge cycles, also abortive apoptosis after extensive chromosomal fragmentation and replication stress have been proposed as potential initiating triggers of chromothripsis [10]. Currently, the most attractive model for chromothripsis

combines both replication stress and mitotic errors with the formation of micronuclei containing mis-segregated anaphase chromosome(s) that undergo defective DNA replication.

The fragmentation created by this first stage of chromothripsis is further repaired by one of the several DNA repair mechanism [11]. So far, both non-replicative repair pathways such as non-homologous end joining (NHEJ) and replication-associated repair pathways such as microhomology-mediated break-induced replication (MMBIR) have been implicated in the reassembly process of pulverized chromosomes [12]. In case of limited sequence overlap, non-homologous end joining has been suggested as the most probable molecular mechanism involved in reconnecting the shattered DNA fragments [3]. In this study, whole genome paired-end sequencing was performed on DNA from the prostate cancer cell line VCaP [13]. This cell line is widely used as a model for research on castration-resistant prostate cancer (CRPC). It is derived from a vertebral metastatic lesion, grows androgen dependently and has an amplified androgen receptor (AR) gene and the common *TMPRSS2-ERG* fusion gene [14]. The genome was sequenced at an average coverage of 113× and an average fully called genome fraction of 97.6 % (Supplementary methods). Overall, we detected 2,414 high confidence structural variation (SV) breakpoint events, of which <4 % juxtaposed portions of gene sequences on both sides (Figure 3.1a). By far, the highest number (573; 24 %) is mapped on chromosome 5q (Figure 3.1b; Table S1). The remarkable clustering of a massive number of rearrangements on the q arm of chromosome 5 is a hallmark of the chromothripsis phenomenon [12]. VCaP has a near-triploid genome, in which most likely the chromothripsis event succeeded the duplication event. We observe, therefore, a copy number distribution varying mainly between three copy number states and maintenance of heterozygosity (Figure 3.2, Figure S1). In addition, the breakpoints in 5q were remarkably clustered in smaller regions with six rearrangements involving a 7-kb region between 81.121 and 81.128 Mb (Table S1), although the normal distance between the two joined fragments is normally tens of megabases.

Out of the 573 SV breakpoints involving 5q, only four were interchromosomal with rearrangements to chromosomes 6 (three events) and 8 (Table S1). Remarkably, sequencing of the breakpoint junctions of these SV breakpoints revealed frequent insertions of novel sequences (186/573) (Table S2). In the gene fusions resulting from complex chromothripsis rearrangements in 5q, these insertions of up to 255 bp corresponded to fragments of chromosome 5 located within 7–53 Mb distance of the adjacent fusion partner. We rarely observe microhomology flanking the breakpoints in our set of gene fusions, and in a few cases, we could detect repeats flanking in both sides of the breakpoint junction (Table S3).

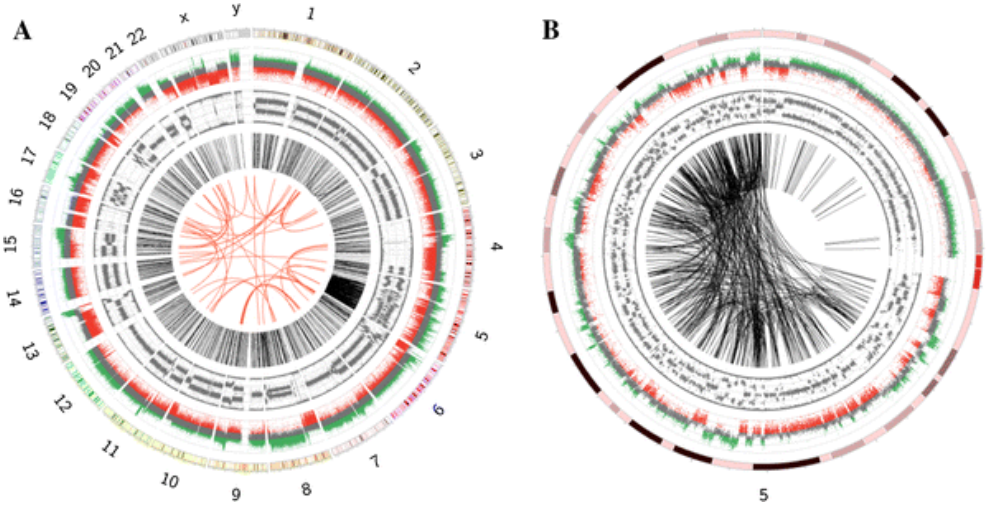


Figure 3.1: Circos plots showing structural and copy number variation across the whole genome (a) and chromosome 5 (b) of VCaP. Each chromosome is represented in the outer ring. The outer data ring corresponds to copy number variation, with regions of gain depicted in green and loss in red. The inner data ring represents B-allele frequency. The intra and interchromosomal rearrangements are on the inside and depicted in black and red lines, respectively

Forty-three out of the 573 SV breakpoint events on 5q were in different genes at both sides. In 18 of the 43 events, the two different genes were in the same orientation and potentially encoding a functional fusion protein (Table 3.1, Figure S2). In order to validate the 18 candidate gene fusions on DNA level and check whether the fusion transcripts encode an in-frame fusion protein, we performed PCR on both DNA and cDNA level. We could validate 15 candidate gene fusions at the DNA level (Figure S3), whereas only one-third were detected on mRNA level suggesting downregulation of gene expression or instability of the fusion transcripts [0]. Sequencing of the PDE4D-FAM172A breakpoint revealed that within this gene fusion a small fragment of PPP2R2B sequence (55 bp) is inserted in-between the PDE4D and FAM172A genes. This explains the correct NGS identification of the PDE4D-PPP2R2B and PPP2R2B-FAM172A breakpoints, but the absence of a PCR product since the PPP2R2B primers were originally designed outside of the small insert. Conversely, the ADAMTS12-PXDNL candidate gene fusion had a very low number of discordant mate pairs indicating the fusion event, and hence is most likely a sequencing artifact. Chromothripsis, being a seemingly random process, results in highly altered chromosomes with numerous mutations and rearrangements. The formation of gene fusions as a consequence of the chromothripsis event does not seem to be preferential over rearrangements occurring outside genes. Moreover, we did not observe positive selection for in-frame fusion transcripts, since only one out of the five expressed fusion transcripts resulted in a feasible fusion protein. The role of this fusion between NDUFAF2 and MAST4 remains to be determined (Figure S4, Table S4).

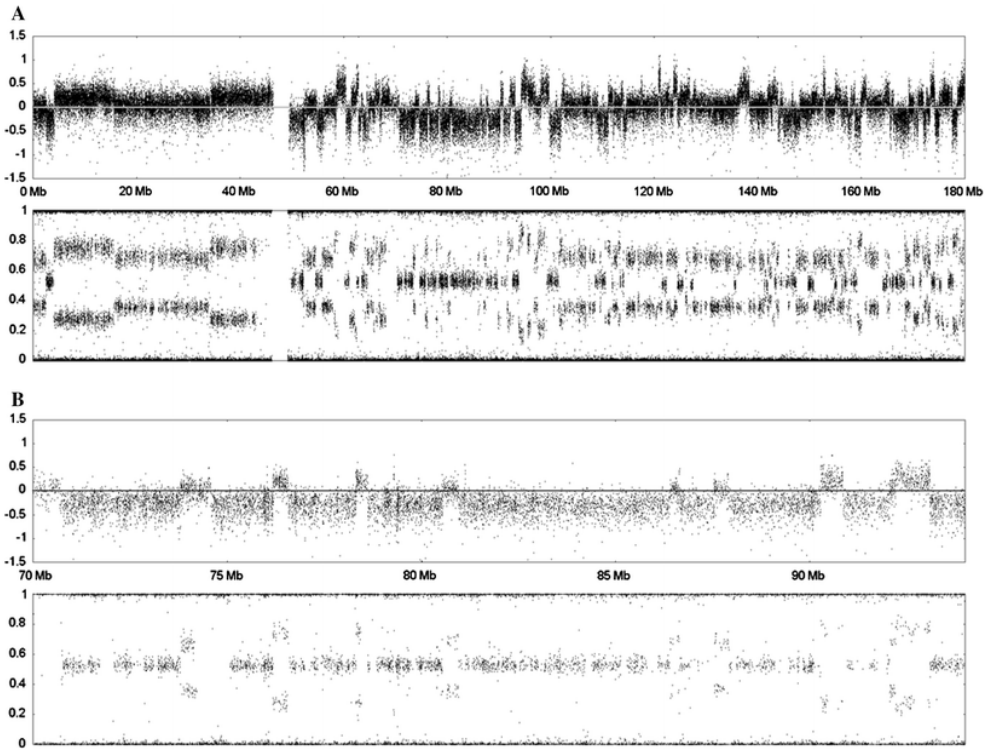


Figure 3.2: Clustered rearrangements on chromosome 5q of VCaP. Copy number across chromosome 5 oscillates between a copy number of 2, 3, and 4. Copy number 2 corresponds to segments of SNP probes below the zero line, copy number 2 to segments of SNP probes in the zero line, and copy number 4 to segments of SNP probes above the zero line. The B-allele frequency plot is displayed below the copy number plot

5' Donor gene	3' Acceptor gene	Donor gene Chr	Acceptor gene Chr	Validated cDNA	Validated DNA	In-frame	# of Discordan mate pairs
PDE4D	C5orf47	5	5	Yes	Yes	No	218
CPLX2	UBXD8	5	5	No	Yes	No	55
EBF1	FBXL17	5	5	No	Yes	No	74
KCNN2	EBF1	5	5	No	Yes	Yes	509
RASGRF2	RNF145	5	5	No	Yes	No	71
JMY	DMGDH	5	5	Yes	Yes	No	102
TRIM40	FBXO38	6	5	No	Yes	No	89
LMAN2	AP3S1	5	5	Yes	Yes	No	191
EFNA5	PCDHB7	5	5	No	Yes	No	11
YTHDC2	PPP2R2B	5	5	No	Yes	No	8
PDE8B	UIMC1	5	5	No	Yes	No	63
ZFP62	RGNEF	5	5	No	Yes	No	225
NDUFAF2	MAST4	5	5	Yes	Yes	Yes	197
ADAMTS12	PXDNL	8	5	No	No	No	3
EBF1	FEM1C	5	5	No	Yes	No	11
PDE4D	FAM172A	5	5	Yes	Yes	No	119
PDE4D	PPP2R2B	5	5	No	No*	No	12
PPP2R2B	FAM172A	5	5	No	No*	Yes	7

Table 3.1: List of gene fusions involving chromosome 5 of VCaP. The column *Donor Gene Chr* refers to the chromosome number of the donor gene, and the column *Acceptor Gene Chr* refers to the chromosome number of the 3' acceptor gene. The column *# of discordant mate pairs* displays the number of mate pair reads that are discordant in relation to the reference genome build 18 and concordant with the respective reported SV event (gene fusion). The * corresponds to by-product events that do not result in a real gene fusion

Using PCR, we observed that all 15 fusions detected in VCaP were also present in the DuCaP cell line which is derived from a dura mater metastasis from the same patient that gave rise to VCaP [15], indicating that chromothripsis occurred in the cells that resulted in both VCaP and DuCaP metastases (Figures S3 and S4) (data not shown). The finding that both VCaP and DuCaP harbor the identical gene rearrangements identified in chromosome 5 indicates that these were present in the precursor prostate adenocarcinoma lesion and not a cell line cultivation artifact.

In our whole genome study of VCaP cells, we also detected a fairly complex rearrangement involving the TMPRSS2 and ERG genes on chromosome 21q. ehe *TMPRSS2-ERG* fusion in VCaP results from the assembly of the ERG and TMPRSS2 breakpoints with the insertion of two fragments of TMPRSS2 (Table S5). This did not disrupt the transcript and open reading frame of the final fusion product [16].

Recently, an association between TP53 mutations and chromothripsis has been observed for acute

myeloid leukemia and pediatric medulloblastoma [3]. In order to determine the status of TP53 in the VCaP cell line, we examined all the single nucleotide variants (SNVs) detected by next generation sequencing in the TP53 gene (Table S6). We observed two homozygous missense SNVs present in the TP53 coding DNA sequence (CDS): c.742C>T and c.215C>G. The c.742C>T (p.R248 W) is a well-known frequently observed mutation that fully inactivates TP53. The c.215C>G (p.R72P) is a natural occurring polymorphism in exon 4 of TP53. As a result of the c.742C>T SNV, the VCaP cell line has no wild type functional p53, a key player in the maintenance of genomic stability. In addition, we have also examined whether there were SNVs present in DNA repair genes previously shown to be altered in prostate cancer [17] (Table S7). Interestingly, we found the genes ATM, MLH1, PRKDC, and ERCC5 to have missense SNVs in their respective CDS. The only SNVs described to be mutational were present both in the ATM gene (COSMIC mutation ID number 21826 and 21827), but the functional relevance of these remains unknown [18]. An SNV present in the intron 7 acceptor splicing site of the XRCC4 gene (rs1805377) has shown to be significantly associated with increased prostate cancer risk [19]. The homozygous missense c.655A>G SNV present in the MLH1 gene (p.I219V) has been associated with increased risk of colorectal cancer [20]. Deficiencies in the DNA mismatch repair system are commonly observed in colorectal cancer and to a less extend in prostate cancer. The influence of the c.655A>G (p.I219V) missense SNV in the development of prostate cancer remains undetermined.

Here, we report the structural variations detected in VCaP and show that the q arm of chromosome 5 has undergone chromothripsis. Chromosome 5 appears to be frequently affected by chromothripsis with studies showing the same pattern in a renal cancer cell line [0] and a neuroblastoma patient [4]. It remains undetermined, however, whether chromosome-specific properties or the nonrandom radial localization of chromosomes in distinct territories play a role in the preferential target of certain chromosomes by chromothripsis. Whether the catastrophic events on 5q in VCaP and the related cell line DuCaP have been playing and still playing an important role in tumorigenesis remains to be determined.

The massive number of genomic breaks occurring in chromosome 5q hampers the formulation of a definitive model for the generation of chromothripsis. The copy number variation and B-allele frequency of chromosome 5q in VCaP is a repetition of regions with n2 (AB), n3 (AAB), and n4 (AAAB) (Figure S1). Based on chromosome painting, VCaP is an overall near-triploid cell line with four copies of chromosome 5 that vary in size (68–74(3n), XXYY) (van Bokhoven et al. 2003). We hypothesize that this pattern is explained by two normal chromosome copies (n2 (AB)) and two copies of A with chromothripsis (n4 (A*A*AB)), of which one of the two copies has undergone additional rearrangements resulting in n3 (A*AB). The sequence of events that would lead to this

pattern is a duplication of A and B (AABB) and subsequent loss of one B-allele (AAB). One of the A-alleles would have undergone chromothripsis (A*AB) and duplication (A*A*AB). Additional rearrangements (deletions or even a second chromothripsis event) would explain the observed 5q regions of n3 (A*AB) (Figure S5). Our study has shown that research on genes located on and transcripts derived from chromosome 5q need to take the chromothripsis into account. Besides, being a widely used model for research on AR, ERG, and CRPC, VCaP might prove a highly relevant model for research on chromothripsis.

Acknowledgements

This study was supported by the Complete Genomics Inc. grant (EMC GL 083111), the FP7 Marie Curie Initial Training Network PRO-NEST (grant number 238278) and the CTMM (Center for Translational Molecular Medicine) Translational Research IT (TraIT).

Bibliography

[0] P. J. Stephens, C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, *et al.*, "Massive genomic rearrangement acquired in a single catastrophic event during cancer development," *cell*, vol. 144, no. 1, pp. 27–40, 2011.

[1] F. Magrangeas, H. Avet-Loiseau, N. C. Munshi, and S. Minvielle, "Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients," *Blood*, vol. 118, no. 3, pp. 675–678, 2011.

[2] W. P. Kloosterman, M. Hoogstraat, O. Paling, M. Tavakoli-Yaraki, I. Renkens, J. S. Vermaat, M. J. van Roosmalen, S. van Lieshout, I. J. Nijman, W. Roessingh, *et al.*, "Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer," *Genome biology*, vol. 12, no. 10, p. R103, 2011.

[3] T. Rausch, D. T. Jones, M. Zapatka, A. M. Stütz, T. Zichner, J. Weischenfeldt, N. Jäger, M. Remke, D. Shih, P. A. Northcott, *et al.*, "Genome sequencing of pediatric medulloblastoma links catastrophic dna rearrangements with tp53 mutations," *Cell*, vol. 148, no. 1, pp. 59–71, 2012.

[4] J. J. Molenaar, J. Koster, D. A. Zwijnenburg, P. van Sluis, L. J. Valentijn, I. van der Ploeg, M. Hamdi, J. van Nes, B. A. Westerman, J. van Arkel, *et al.*, "Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes," *Nature*, vol. 483, no. 7391, p. 589, 2012.

[5] C. A. Maher and R. K. Wilson, "Chromothripsis and human disease: piecing together the shattering process," *Cell*, vol. 148, no. 1-2, pp. 29–32, 2012.

[6] P. J. Campbell, P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li, T. Santarius, L. A. Stebbings, C. Leroy, S. Edkins, C. Hardy, *et al.*, "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing," *Nature genetics*, vol. 40, no. 6, p. 722, 2008.

[7] P. J. Stephens, D. J. McBride, M.-L. Lin, I. Varela, E. D. Pleasance, J. T. Simpson, L. A. Stebbings, C. Leroy, S. Edkins, L. J. Mudie, *et al.*, "Complex landscapes of somatic rearrangement in human breast cancer genomes," *Nature*, vol. 462, no. 7276, p. 1005, 2009.

[8] C. Righolt and S. Mai, "Shattered and stitched chromosomes—chromothripsis and chromoanasynthesis—manifestations of a new chromosome crisis?," *Genes, Chromosomes and Cancer*, vol. 51, no. 11, pp. 975–981, 2012.

[9] T. Misteli, "Beyond the sequence: cellular organization of genome function," *Cell*, vol. 128, no. 4, pp. 787–800, 2007.

[10] M. J. Jones and P. V. Jallepalli, "Chromothripsis: chromosomes in crisis," *Developmental cell*, vol. 23, no. 5, pp. 908–917, 2012.

[11] A. J. Holland and D. W. Cleveland, "Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements," *Nature medicine*, vol. 18, no. 11, p. 1630, 2012.

[12] J. V. Forment, A. Kaidi, and S. P. Jackson, "Chromothripsis and cancer: causes and consequences of chromosome shattering," *Nature Reviews Cancer*, vol. 12, no. 10, p. 663, 2012.

[13] R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, *et al.*, "Human genome sequencing using unchained base reads on self-assembling dna nanoarrays," *Science*, vol. 327, no. 5961, pp. 78–81, 2010.

[14] S. Korenchuk, J. Lehr, L. MClean, Y. Lee, S. Whitney, R. Vessella, D. Lin, and K. Pienta, "Vcap, a cell-based model system of human prostate cancer.," *In vivo (Athens, Greece)*, vol. 15, no. 2, pp. 163–168, 2001.

[15] Y. Lee, S. Korenchuk, J. Lehr, S. Whitney, R. Vessela, and K. Pienta, "Establishment and characterization of a new human prostatic cancer cell line: Ducap.," *In vivo (Athens, Greece)*, vol. 15, no. 2, pp. 157–162, 2001.

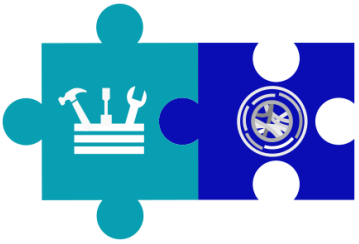
[16] K. D. Mertz, S. R. Setlur, S. M. Dhanasekaran, F. Demichelis, S. Perner, S. Tomlins, J. Tchinda, B. Laxman, R. L. Vessella, R. Beroukhimt, *et al.*, "Molecular characterization of tmprss2-erg gene fusion in the nci-h660 prostate cancer cell line: a new perspective for an old model," *Neoplasia*, vol. 9, no. 3, pp. 200–206, 2007.

[17] C. S. Grasso, Y.-M. Wu, D. R. Robinson, X. Cao, S. M. Dhanasekaran, A. P. Khan, M. J. Quist, X. Jing, R. J. Lonigro, J. C. Brenner, *et al.*, "The mutational landscape of lethal castration-resistant prostate cancer," *Nature*, vol. 487, no. 7406, p. 239, 2012.

[18] F. Gumy-Pause, P. Wacker, P. Maillet, D. R. Betts, and A.-P. Sappino, "Atm alterations in childhood non-hodgkin lymphoma," *Cancer genetics and cytogenetics*, vol. 166, no. 2, pp. 101–111, 2006.

[19] R. K. Mandal, V. Singh, R. Kapoor, and R. Mittal, "Do polymorphisms in xrcc4 influence prostate cancer susceptibility in north indian population?," *Biomarkers*, vol. 16, no. 3, pp. 236–242, 2011.

[20] P. T. Campbell, K. Curtin, C. M. Ulrich, W. S. Samowitz, J. Bigler, C. M. Velicer, B. Caan, J. D. Potter, and M. L. Slattery, "Mismatch repair polymorphisms and risk of colon cancer, tumour microsatellite instability and interactions with lifestyle factors," *Gut*, vol. 58, no. 5, pp. 661–667, 2009.



"I didn't want to just know names of things. I remember really wanting to know how it all worked."

Elizabeth Blackburn

4

Somatic Variant Detection

Sequencing of tumours often involves the co-sequencing of a sample of healthy tissue from the same individual. This allows us to distinguish the tumour-specific variants from those present in the germline of the patient. When such a matching normal sample is not available, the analysis of variants becomes more complex. In this chapter, we investigate whether a *virtual normal* may be used in lieu of an associated normal in such cases. A virtual normal consists of a large set of samples from healthy, unrelated, genetically diverse individuals.

As a first step in this analysis, we integrated a suite of tools for variant analysis and visualisation into Galaxy (CGtag). We then used these tools to compare performance of a virtual normal as an alternative to a matched tumour-normal pair.

This chapter contains the following sub-chapters:

4.0 **CGtag: Complete Genomics Toolkit and Annotation in a Cloud-based Galaxy** Different sequencing techniques are constantly being developed, and often come with custom analysis tools. This was also the case for Complete Genomics (CG) data. CG defined it's own data format and supplied specialized analysis tools, optimized for their own sequencing data. These were command-line tools, and in order to increase their accessibility, I *wrapped* them (made available) in Galaxy. Furthermore, to increase interoperability of the CG tools and datasets, I created some additional tools for file format conversion, to allow other, non-CG tools to be applied to CG datasets. Note that while this publication mentions availability in the CTMM-TraIT Tool Shed, the tools have since moved to the main Galaxy Tool Shed. While the tools are still available for analysis or historic data, the tools have become obsolete as the sequencing techniques have evolved.

4.1 **Discriminating somatic and germline mutations in tumor DNA samples without matching normals**. In this work, I used CG-sequenced tumour-normal pairs and CG analysis tools to evaluate the utility of using as large set of samples from diverse, healthy and unrelated individuals, to serve as a *virtual normal sample*, in absence of the commonly used *associated normal* sample originating from healthy tissue from the same patient. As with the previous sub-chapter, the methods and code developed for this work are still available (from the main Galaxy Tool Shed rather than the CTMM-TraIT Galaxy), but have become somewhat obsolete now, since they were geared specifically towards CG-datasets. The approach is still valid, but would require some re-implementation to work for today's sequencing data. Such is often the nature of modern bioinformatics; the high speed of technological advances in the sequencing techniques leaves a high turnover of analysis tools in its wake.

CGtag: Complete Genomics Toolkit and Annotation in a Cloud-based Galaxy

Saskia Hiltemann[1,*], Hailiang Mei[2], Mattias de Hollander[3], Ivo Palli[1], Peter van der Spek[1], Guido Jenster[4], Andrew Stubbs[3]

1. Department of Bioinformatics, Erasmus Medical Centre, Rotterdam, The Netherlands.
2. Netherlands Bioinformatics Center, NBIC, Nijmegen, The Netherlands
3. Department of Microbial Ecology, Netherlands Institute of Ecology, NIOO-KNAW, Wageningen, The Netherlands
4. Department of Urology, Erasmus Medical Centre, Rotterdam, The Netherlands

Published in: *GigaScience*, 2014 Jan 24;3(1):1
DOI: https://doi.org/10.1186/2047-217X-3-1

ABSTRACT

Background: Complete Genomics provides an open-source suite of command line tools for the analysis of their CG-formatted mapped sequencing files. Determination of, for example, the functional impact of detected variants, requires annotation with various databases that often require command line and/or programming experience; thus, limiting their use to the average research scientist. We have therefore implemented this CG toolkit, together with a number of annotation, visualisation and file manipulation tools in Galaxy called CGtag (Complete Genomics Toolkit and Annotation in a Cloud-based Galaxy).

Findings: In order to provide research scientists with web-based, simple and accurate analytical and visualisation applications for the selection of candidate mutations from Complete Genomics

data, we have implemented the open-source Complete Genomics tool set, CGATools, in Galaxy. In addition we implemented some of the most popular command line annotation and visualisation tools to allow research scientists to select candidate pathological mutations (SNV, and indels). Furthermore, we have developed a cloud-based public Galaxy instance to host the CGtag toolkit and other associated modules.

Conclusions: CGtag provides a user-friendly interface to all research scientists wishing to select candidate variants from CG or other next-generation sequencing platforms' data. By using a cloud-based infrastructure, we can also assure sufficient and on-demand computation and storage resources to handle the analysis tasks. The tools are freely available for use from an NBIC/CTMM-TraIT (The Netherlands Bioinformatics Center/Center for Translational Molecular Medicine) cloud-based Galaxy instance, or can be installed to a local (production) Galaxy via the NBIC Galaxy tool shed.

Keywords: Complete Genomics, Next Generation Sequencing, Genetic variation, Pathogenic gene selection.

Findings

Background

Complete Genomics (CG) supplies results for whole-genome next-generation sequencing (NGS) data mapped to a user-defined genome [0] and additional open-source tools [1] for further characterisation of the sequenced genomes. Whilst these tools are open-source and available for download and use on the command line, they are not amenable for scientists to use from their desktop, and require scripting skills to link these tools together with other applications to successfully prioritise candidate pathogenic genes based on these NGS results. To address this issue, we implemented the Complete Genomics Analysis Toolkit (CGATools), including several functional annotation and visualisation tools in a cloud-enabled instance of Galaxy. Galaxy offers a web-based graphical user interface to command line tools, and allows for the graphical construction of complex workflows; Galaxy will automatically keep track of the analysis history, and allows for easy sharing and publishing of data and/or workflows with other users [2, 3, 4]. Furthermore, Galaxy is an extensible platform, nearly any software tool may be integrated into Galaxy, and there is an active community of users and developers ensuring the latest tools are made available for use in Galaxy through the Galaxy tool shed.

This implementation of the CGATools in a Galaxy environment simplifies the analysis of genomes

via the Galaxy GUI and the cloud resource ensures that sufficient computing power is available for the analysis. The inherent functionality in Galaxy of CGtag enables the creation of customisable user-defined workflows by the scientist and not only by the bioinformatician.

For large datasets, transfer to Galaxy via SFTP is available and recommended, but is still limited by the upload speed of the user's internet connection, and can be a bottleneck in the analysis of large datasets.

Variant Detection

CGATools is an open-source project to provide tools for downstream analysis of Complete Genomics data, and may be downloaded from their repository [1]. These tools must be run from the command line and are therefore, not accessible to all users. To remedy this, Complete Genomics also provide Galaxy tool wrappers for many of the CGAtools, which can be downloaded from the Main Galaxy tool repository (tool shed) [5]. However, these Galaxy tools still need to be installed on the users' local (production) Galaxy instance before they can be utilised. We have now made these tools available on a public server [6], and have added Galaxy wrappers for those CGAtools that were not provided by Complete Genomics e.g. Junctions2Events, makeVCF (Table 4.0). The use of the CGAtools in Table 4.0 have previously been outlined [7], using a combination of ListVariants and TestVariants or CallDiff to determine candidate pathogenic single nucleotide variants (SNVs), indels and subs in a selected genome as compared with on or more reference genomes or as part of a trio based genetic analysis [7]. The VarFilter may be used to select those variants which have a high confidence based on the underlying sequence reads as specified as VQHIGH, and the SNPDiff tool can then be used to determine concordance of the NGS results with those of an orthogonal SNV detection platform such as an Affymetrix or Illumina SNP array. The JunctionDiff and Junction2Events tools are used to select fusion events and candidate fusion genes based on quality of the discordant reads used to detect the structural variation event [8].

Functional Annotation Tools

To provide users with enhanced filtering capabilities, we have integrated several command line annotation tools in this NBIC/CTMM-TraIT Galaxy instance. ANNOVAR [9] is a command line tool used to functionally annotate genetic variants. We provide a Galaxy tool wrapper for ANNOVAR. This tool will take a list of variants as input and provide gene and amino acid change annotation, SIFT scores, PolyPhen scores, LRT scores, MutationTaster scores, PhyloP conservation scores, GERP++ conservation scores, DGV variant annotation, dbSNP identifiers, 1000 Genomes Project allele frequencies, NHLBI-ESP 6500 exome project allele frequencies, and

Function	CGATool	Description
Variant detection	CGATools ListVariants	Lists the non-redundant set of small variations found in an arbitrary number of genomes.
Variant detection	CGATools TestVariants	Determine which variants are found in which genomes given the results of ListVariants.
Variant detection	CGATools CallDiff	Compares two variant files to determine where and how the genomes differ.
Variant detection	CGATools VarFilter	Copies the input varfile or masterVar file, applying filters.
Variant detection	CGATools JunctionDiff	Reports difference between junction calls of CG junction files.
Variant detection	CGATools Junctions2Events	Groups and annotate related junctions.
Quality control	CGATools SNPDiff	Compares genotype calls to CG variant files.
File merge	CGATools Join	Merge two tab-delimited files based on equal field or overlapping regions.
Sequence retrieval	CGATools DecodeCRR	Retrieve sequences from a CRR file for a given range of a chromosome.
File conversion	CGATools mkVCF	Converts CG variant and/or junction files to VCF.
File conversion	CGATools generateMasterVar	Converts a varfile to a on-line-per-locus format.
File conversion	CGATools fasta-2-crr	Converts fasta sequences into a single reference crr file.
File conversion	CGATools crr-2-fasta	Converts crr file to fasta sequence.
File conversion	TestVariants2VCF	CG community tool. Converts output of the TestVariants tool to VCF.
Annotation	ANNOVAR	Functional annotation of genetic variants from high-throughput sequencing data.
Annotation	MutationAssessor	Functional impact of protein mutations.
Annotation	Condel	CONsensus DELeteriousness score of missense SNVs.
Visualisation	CG Circos plots	Create CG-style tumour, normal and somatic plots.
Visualisation	Integrative plot	Create circos plot from CG and SNParray data.
Visualisation	Generic genomic data plotter	Plot any type of numerical genomic data using GNUplot.
File manipulation	Filter columns	Filter tab-delimited files based on column contents.
File manipulation	Add/remove chr prefix	adds or removes chr prefix from chromosome column.
File manipulation	Column select	Extract and/or rearrange columns in tab-delimited file.
File manipulation	Sort chromosomal position	Sort a tab-delimited file by chromosomal position.
File manipulation	Strip header	Remove header from files.
File manipulation	File concatenation	Concatenate 2 files (e.g. for restoring header).

Table 4.0: Overview of CGTag tools available in NBIC/CTMM-TraIT Galaxy and the NBIC tool shed

other information. We have implemented this tool to accept VCF (v4) files, Complete Genomics varfiles or CG-derived tab-separated files using the CG 0-based half-open coordinate system, or lastly, the standard ANNOVAR input format consisting of tab-separated lists of variants using the 1-based coordinate system. This tool will output the original file columns, followed by additional ANNOVAR columns. The ANNOVAR code itself is not included in the tool shed repository, but instructions on how to obtain a license and the subsequent manual installation of the tool are included in the readme of the Galaxy tool shed repository. We obtained permission to offer ANNOVAR on our public Galaxy server, so the tool can be previewed there. To supplement

ANNOVAR, Condel (CONsensus DELeteriousness) [10] has been included to calculate the deleterious score associated of missense SNVs and the impact of non-synonymous SNVs on protein function. Condel integrates the outputs of two tools: SIFT and Polyphen2, to calculate a weighted average of the scores (WAS) of these tools. Condel can optionally incorporate the output of a third tool, MutationAssessor, which is also included in this Galaxy instance. Mutation Assessor [11] is a web-based tool providing predictions of the functional impact of amino-acid substitutions in proteins, such as mutations discovered in cancer or missense polymorphisms. The MutationAssessor database is accessed through a REST API. In order not to overload the server, queries are limited to 3 per second, so when dealing with a long list of variants, some pre-filtering is recommended. The functional annotation provided by ANNOVAR, including the addition of multiple versions of dbSNP, the variants provided by Complete Genomics Public data from unrelated individuals only [12] and 31 genomes from Huvariome [13], are available in this Galaxy instance. Huvariome provides the user with additional whole genome variant calls for those regions which are difficult to sequence and can retrieve the weighted allele frequency for each base in the human genome [13].

Visualisation Tools

A generic genomic data plotter tool based on GNUplot is available, which takes as input, a tab-delimited file of format chr–start-end–value, and will output either a single chromosome plot, an overview of all chromosome plots in a single image, or a sub-region of a chromosome defined by the user. Additionally, the tool has the option of plotting input from a second file in the same image, which is useful for tumour-normal comparison (Figure 4.0). B-allele frequency (BAF) is used to determine whether the structural variation junction is homo- or heterozygous. When the data is in the right format, the generic plotter tool can be used to visualise the BAF, and we have also implemented a plot tool to display allele frequencies directly from a CG masterVar file, again with the capability of displaying single-chromosome plots, all chromosomes in a single image, or custom defined regions (Figure 4.0). The current Complete Genomics analysis pipeline (CGAP v2.5) delivers Circos [14] visualisations with each genome that is sequenced and the code used to generate these images have been made freely available for download [15]. We have modified this code and implemented Galaxy tools to allow for the generation of these images for samples sequenced on earlier CG analysis pipelines (before v2.0), that utilise the junctions file, masterVar file, CNV details and CNV segments files to generate the standard CG Circos report.

To support fusion gene analysis we have created a custom Circos tool which uses CG files, CG junctions file and CG varfile for NGS, and the results from SNP arrays analysis, specifically

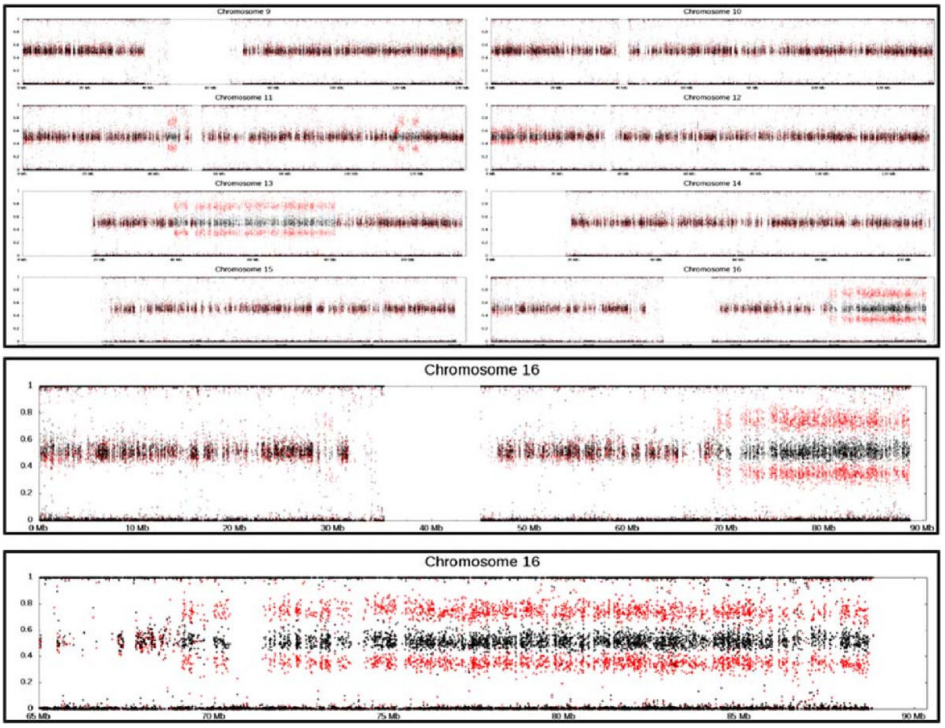


Figure 4.0: Generic Genomic data plotting tool. Output from our generic genomic data plotter used to plot B-allele frequency from Illumina 1M SNParray data. Plot with two tracks; tumour (red) and normal (black). Output can be (top) a whole genome overview (shown here in part), or (middle) a single chromosome, or (bottom) a subregion of a chromosome defined by the user (here chr16, 60MB-end). Many parameters such as the colour and sizes of the data points may be adjusted by the user as required.

the B-allele frequency (BAF) and copy number variation (CNV) files. The output is either a whole-genome plot, per-chromosome plots, a single image containing all the per-chromosome plots together, or a plot of a custom region defined by the user (e.g., a plot showing just chromosomes 3, 5, and X, or a plot showing a specific range within a single chromosome). Additionally the user can select an "impacted genes" track for the per-chromosome plots, which will print the names of the genes impacted by SV events along the outer edge of the image (Figure 4.1). This custom Circos script is capable of using fusion gene detection results generated from the Illumina platform with the fusion genes detected by an application such as FusionMap [16], and which are reported in custom FusionMap report format, a tab-delimited file similar to that delivered by Complete Genomics.

In addition to these tools within Galaxy, structural variation files processed using CGtag may be exported to our previously described fusion gene prioritisation tool, iFUSE [17] to identify

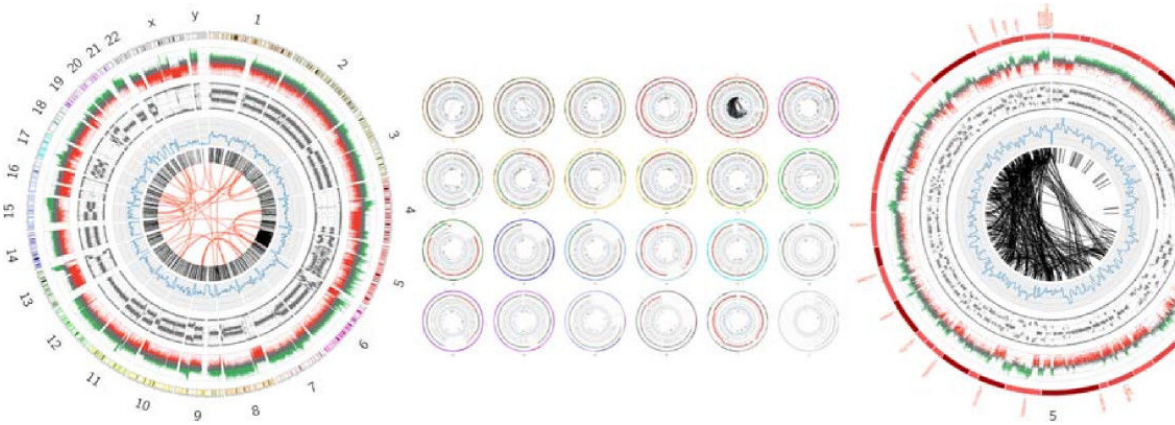


Figure 4.1: Circos Integrative Plot Tool. Circos plots for (left) whole genome, (middle) overview or all chromosomes in single images, and (right) for a single chromosome. Each chromosome is represented in the outer ring and then from outer to inner rings represent copy number variation (with regions of gain depicted in green and loss in red), B-allele frequency, SNP density and the intra- and interchromosomal rearrangements are on the inside and depicted in black and red lines, respectively. Impacted genes track (red gene symbols) are displayed outside the outer chromosome ring and only on the single chromosome plot.

candidate fusion genes and display their representative DNA, RNA and protein sequence.

Auxiliary Tools

Our suite of tools also includes several auxiliary tools supplied by CG but not available from the Galaxy tool shed which offer the user several file format conversion tools (Table 4.0) that enable users to connect the output from the CGATools analysis to other analytical or annotation workflows by means of standard file formats (e.g., FASTA, VCF). In addition a number of file formatting tools are also included, such as removing of headers from files (required by some tools), adding removing of a chr prefix to a column of a file (i.e., chrX vs. X), concatenation of files, and extracting and rearranging of columns, to help facilitate the flow of data from one tool to the next.

CLOUD Implementation

NBIC Galaxy is hosted at a high performance computing (HPC) cloud system operated by SURFsara [18]. This HPC cloud consists of 19 fast servers with 608 CPUs and almost 5TB of memory. The NBIC Galaxy that operates in this HPC cloud is implemented using the Cloudman framework [19] and its adapted version supports the OpenNebula Cloud environment. The advantage of using the Cloudman framework to build NBIC Galaxy is mainly two-fold, firstly Cloudman provides a set of complete scripts to automatically install tools and datasets on a virtual machine image. The installed tools include the Galaxy system itself and all its dependencies. These

dependencies include webserver (nginx), database (postgres), cluster job scheduler (SGE), and common NGS tools, such as bowtie, BWA, samtools, and so forth. The installed datasets include most of the common reference genomes (hg18, hg19, mm9, etc) and their tool-specific index files. Thus, the end product of running Cloudman installation script is a fully functional NBIC Galaxy system operating in the HPC Cloud.

The second contribution of Cloudman to our NBIC Galaxy system is its ability to set up a flexible virtual cluster and ability to provide auto-scaling support. The previous NBIC Galaxy was hosted on a dedicate physical server with rather limit resources (4 CPU, 32G memory). Due to this resource limitation, our NBIC Galaxy was never promoted to be a real data analysis server to handle the production level of NGS datasets. On the other hand, because of the sporadic nature of user access, the server was mostly on idle during its 2-year lifespan. Moving to Cloud resolved both issues. The current NBIC Galaxy operates on top of a virtual cluster. This virtual cluster contains one head node and a number of worker nodes. These nodes are all virtual machines that are built using the machine image generated by the Cloudman script. During minimal usage, the cluster will only contain one head node. Once a significant load occurs due to training courses or production level data analysis, the virtual cluster can automatically scale itself upwards. More worker nodes will be added dynamically to this virtual cluster to boost the capacity of NBIC Galaxy. Once the load decreases, the virtual cluster can scale down again to operate with only a limited number of nodes.

The use of shared resources does have drawback as well. We have experienced a more obvious I/O bottleneck in the cloud-based NBIC Galaxy compared to the previous system that ran in a physical machine. In the HPC Cloud, storage is provided through a network file system (NFS) instead of a local hard disk. When more concurrent Cloud users are using the Cloud resource, we observe the extra job time caused by I/O delays. However, we argue that this issue is far outweighed by the benefit of having a dynamic virtual cluster support to the NBIC Galaxy.

Availability and requirments

Project Name: CGtag: Complete Genomics Toolkit and Annotation in a Cloud-based Galaxy
Project home page: http://galaxy-demo.trait-ctmm.cloudlet.sara.nl
Operating system: Linux (Galaxy and CGtag)
Programming language: Python (Galaxy and CGtag), R (CGtag), Bash (CGTag)
Other requirements: Circos [14], GNUplot [20], Complete Genomics open source Toolset [1] and dependencies therein; see documentation for a comprehensive list of optional dependencies, based on workflow requirements.

License: GPL v3
Any restrictions to use by non-academics: ANNOVAR license must be obtained before it can be used
published page: http://galaxy.ctmm-trait.nl/u/saskia-hiltemann/p/cgtag
Links to tool shed repositories:
annovar: http://toolshed.nbic.nl/view/saskia-hiltemann/annovar
cgatools: http://toolshed.nbic.nl/view/saskia-hiltemann/cgatools_v17
circos plotters: http://toolshed.nbic.nl/view/saskia-hiltemann/cg_circos_plots
condel: http://toolshed.nbic.nl/view/saskia-hiltemann/condel
file manipulation tools: http://toolshed.nbic.nl/view/saskia-hiltemann/file_manipulation
generic genomic data plotter: http://toolshed.nbic.nl/view/saskia-hiltemann/genomic_data_plotter
mutation assessor: http://toolshed.nbic.nl/view/saskia-hiltemann/mutation_assessor
NOTE: these tools can be installed to both Cloudman Galaxy instances or non-Cloudman Galaxy instances alike (via the tool shed or manually from the command line).

Availability and supporting data

All tools described, as well as example data, are available from the NBIC/CTMM-TraIT Galaxy server (http://galaxy.ctmm-trait.nl) and the NBIC Galaxy tool shed (http://toolshed.nbic.nl).

Abbreviations

BAF: B-allele Frequency
CG: Complete Genomics
CGATools: Complete Genomics analysis tools
CGtag: Complete Genomics Toolkit and Annotation in a Cloud-based Galaxy
NBIC: The Netherlands Bioinformatics Center
NFS: Network File System
SNV: Single Nucleotide Variation
SV: Structural Variation

Declarations

Acknowledgements

This study was performed within the framework of CTMM, the Center for Translational Molecular Medicine. TraIT project (grant 05T-401).

This work was sponsored by the BiG Grid project for the use of the computing and storage facilities, with financial support from the Netherlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organisation for Scientific Research, NWO)

Note from the Editors

This paper is part of the GigaScience Galaxy series. We will be hosting some of the computational resources of these papers on our GigaGalaxy server (http://galaxy.cbiit.cuhk.edu.hk).

Competing interests

The authors declare that they have no competing interests.

Authors contributions

SH, GJ, HM and AS contributed to the design and coordination of CGtag and manuscript preparation. SH, MdH, IP and HM contributed to implementing CGtag. SH, GJ, PvdS and AS contributed to testing of CGtag. MdH and HM implemented Galaxy Cloudman for the SURFsara/Big-grid HPC cloud. All authors read and approved the final manuscript.

Bibliography

[0] Y. Ma, S. Dobbins, A. Sherborne, D. Chubb, M. Galbiati, G. Cazzaniga, C. Micalizzi, R. Tearle, A. Lloyd, R. Hain, M. Greaves, and R. Houlston, "Developmental timing of mutations revealed by whole-genome sequencing of twins with acute lymphoblastic leukemia," *Proc Natl Acad Sci USA*, vol. 110, pp. 7429–7433, April 2013.

[1] "Cgatools." http://cgatools.sourceforge.net/.

[2] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome biology*, vol. 11, no. 8, p. R86, 2010.

[3] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor, "Galaxy: A web-based genome analysis tool for experimentalists," *Curr Protoc in Mol Biol*, pp. 19.10.1 – –19.10.21, 2010.

[4] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko, "Galaxy: a platform for interactive large-scale genome analysis," *Genome Res*, vol. 15, no. 10, pp. 1451–1455, 2005.

[5] "Main galaxy toolshed." http://toolshed.g2.bx.psu.edu/.

[6] "Nbic/ctmm-trait cloud-based galaxy." [http://galaxy-demo.trait-ctmm.cloudlet.sara.nl.

[7] P. Nieminen, N. V. Morgan, A. L. Fenwick, S. Parmanen, L. Veistinen, M. L. Mikkola, P. J. van der Spek, A. Giraud, L. Judd, S. Arte, L. A. Brueton, S. A. Wall, I. M. Mathijssen, A. O. Maher, E R amd Wilkie, S. Kreiborg, and I. Thesleff, "Inactivation of il11 signaling causes craniosynostosis, delayed tooth eruption, and supernumerary teeth," *Am J Hum Genet*, vol. 89, pp. 67–81, July 2011.

[8] S. Hiltemann, E. McClellan, J. van Nijnatten, S. Horsman, I. Palli, I. Teles Alves, T. Hartjes, J. Trapman, P. van der Spek, G. Jenster, and A. Stubbs, "ifuse: integrated fusion gene explorer," *Bioinformatics*, vol. 29, pp. 1700–1701, July 2013.

[9] K. Wang, M. Li, and H. Hakonarson, "Annovar: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Res*, vol. 38, p. e164, September 2010.

[10] A. González-Pérez and N. López-Bigas, "Improving the assessment of the outcome of nonsynonymous snvs with a consensus deleteriousness score, condel," *Am J Hum Genet*, vol. 88, pp. 440–449, April 2011.

[11] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: application to cancer genomics," *Nucleic Acids Res*, vol. 39, p. e118, September 2011.

[12] "Complete genomics ftp." ftp.completegenomics.com,ftp2.completegenomics.com.

[13] A. Stubbs, E. McClellan, S. Horsman, S. Hiltemann, I. Palli, S. Nouwens, A. Koning, F. Hoogland, J. Reumers, D. Heijsman, S. Swagemakers, A. Kremer, J. Meijerink, D. Lambrechts, and P. van der Spek, "Huvariome: a web server resource of whole genome next-generation sequencing allelic frequencies to aid in pathological candidate gene selection," *J Clin Bioinformatics*, vol. 2, p. 19, November 2012.

[14] "Circos circular visualisation." http://circos.ca.

[15] "Cg circos scripts." ftp://ftp.completegenomics.com/ToolRepository/CompleteCircosPackage.zip.

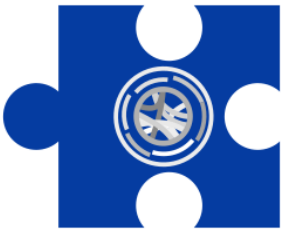
[16] H. Ge, K. Liu, T. Juan, F. Fang, M. Newman, and W. Hoeck, "Fusionmap: detecting fusion genes from next-generation sequencing data at base-pair resolution," *Bioinformatics*, vol. 27, no. 14, pp. 1922–1928, 2011.

[17] "ifuse: integrated fusion gene explorer." http://ifuse.erasmusmc.nl.

[18] "Surf sara hpc cloud." http://www.surfsara.nl/systems/hpc-cloud.

[19] E. Afgan, D. Baker, N. Coraor, B. Chapman, A. Nekrutenko, and J. Taylor, "Galaxy cloudman: delivering cloud compute clusters," *BMC Bioinformatics*, vol. 11, p. S4, December 2010.

[20] "Gnuplot." http://www.gnuplot.info.



Discriminating somatic and germline mutations in tumor DNA samples without matching normals

Saskia Hiltemann[1], Guido Jenster[2], Jan Trapman[3], Peter van der Spek[1], Andrew Stubbs[1]

1. Department of Bioinformatics, Erasmus Medical Center, Rotterdam, The Netherlands.
2. Department of Urology, Erasmus Medical Center, Rotterdam, The Netherlands
3. Department of Pathology, Erasmus Medical Center, Rotterdam, The Netherlands.

Published in: *Genome Research*, 2015 Sep; 25(9): 1382–1390
DOI: 10.1101/gr.183053.114

Abstract

Tumor analyses commonly employ a correction with a matched normal (MN), a sample from healthy tissue of the same individual, in order to distinguish germline mutations from somatic mutations. Since the majority of variants found in an individual are thought to be common within the population, we constructed a set of 931 samples from healthy, unrelated individuals, originating from two different sequencing platforms, to serve as a virtual normal (VN) in the absence of such an associated normal sample. Our approach removed (1) >96% of the germline variants also removed by the MN sample and (2) a large number (2%–8%) of additional variants not corrected for by the associated normal. The combination of the VN with the MN improved the correction for polymorphisms significantly, with up to ~30% compared with MN and ~15% compared with VN only. We determined the number of unrelated genomes needed in order to correct at least as efficiently as the MN is about 200 for structural variations (SVs) and about 400 for single-nucleotide variants (SNVs) and indels. In addition, we propose that the removal of common

variants with purely position-based methods is inaccurate and incurs additional false-positive somatic variants, and more sophisticated algorithms, which are capable of leveraging information about the area surrounding variants, are needed for optimal accuracy. Our VN correction method can be used to analyze any list of variants, regardless of sequencing platform of origin. This VN methodology is available for use on our public Galaxy server.

Introduction

Analysis of 1092 human genomes performed by the 1000 Genomes Project reveals that an individual has approximately 4 million variations (on average, 3.7 million SNPs, 350,000 insertions and deletions [indels], and 750 large deletions) compared with the reference genome and that the vast majority of an individual's germline variations are polymorphic within the human population, with >95% of all single-nucleotide variants (SNVs) and small indels in a given individual occurring at a frequency of ≥0.5% [0, 1]. Therefore, whenever a matched normal (MN) sample was unavailable (most commonly due to lack of funds or sample availability), researchers have typically relied on the public mutation databases and/or a set of in-house genomes for the filtering of germline variants from the full set of variants found in a tumor sample [2, 3]. In recent years, these catalogs of human variation have grown exponentially, causing some researchers to question the necessity of sequencing a MN control for every tumor sample [3].

In this study, we address the questions of whether current mutation databases are complete enough to correct for common and rare polymorphisms and of how well this filtering performs compared with the correction with a MN sample.

There are many public databases of human variation available. The Single Nucleotide Polymorphism Database (dbSNP) is a free public archive for genetic variation within and across different species [4]. Its latest build (138) contains over 63 million polymorphisms found within the human population. The 1000 Genomes Project (1000G) database contains polymorphisms encountered in a set of 1092 genomes of healthy individuals [0, 1]. The NHLBI Exome Variant Server (EVS) contains exonic variants from over 6500 genomes (`http://evs.gs.washington.edu/EVS`).

In an effort to improve the control-free correction method further, we constructed what we call a virtual normal (VN). This is a set of 931 samples from healthy, unrelated individuals, whole-genome sequenced to high depth, originating from two different sequencing platforms. Our VN consists of 433 public samples from Complete Genomics [5], sequenced in the context of the 1000G, as well as

498 samples sequenced on Illumina HiSeq technology by the Genome of the Netherlands (GoNL) Consortium [6, 7].

For copy-number analysis of sequencing data, tools exist that correct for normal contamination in unmatched tumor samples [8]. The idea of using a set of genomes for correction of copy-number variants has also been described [2]. Apart from a correction based on GC content, this read-depth method also corrects for regions found to have an increased or decreased copy number across all five of their samples (from healthy individuals) and therefore likely a polymorphism within the population. We aim to assess the validity of such an approach and extend it by applying it to structural variation (SV), as well as SNV and indel analysis, in whole-genome-sequenced cancer samples. Additionally, we investigate the minimal size of such a VN necessary for adequate filtering and assess the influence of different ethnicities within the set.

Variants in mutation databases are usually represented only by their position relative to the reference genome and the variant allele, as well as possibly a quality metric. The advantage of using a VN is that we can also leverage information about the area surrounding a variant (e.g., nearby variants in the same sample) to optimize correction. There are often several different ways to describe the same variant, possibly involving (slightly) different chromosomal positions, which means that comparison methods that require an exact match of position and variant nucleotides may be suboptimal, leading to false-positive somatic variants. The algorithm we use with our VN correction is capable of detecting equivalences of differently described variants by taking into account the reference sequence surrounding a variant, as well as neighboring variants. This provides a valuable improvement over correction using variant databases, where this contextual information is lost.

Results

We evaluate the performance of the different correction methods on four tumor-normal pairs from two different tumor types (breast and prostate cancer). All of the samples were sequenced by Complete Genomics. Two of the samples were also sequenced using Illumina technology. We evaluate the correction of SNVs and indels of up to ~50 bp, as well as the larger SVs. For the breast cancer samples, additional validation data were available from the COSMIC database, and we used these confirmed variants to assess the performance of the three different correction methods (MN, mutation databases, VN).

We have made our VN correction method available as a tool for the Galaxy workflow platform [9,

10, 11] as part of our tumor analysis in Galaxy (TAG) tool suite [12].

MN vs. VN

We use three correction methods in order to determine the set of tumor-specific (somatic) variants: a correction for germline variants using a MN sample, a correction for polymorphisms using the VN, and a correction for polymorphisms using public mutation databases (dbSNP, 1000G, and EVS). All variants remaining after application of all three correction methods represent the (consensus) set of true somatic variants. A false-positive variant of a method is any variant remaining after correction, which would have been removed had we employed all three correction methods.

Structural variations

There are far fewer common SVs, than SNVs and indels, present in public databases. And the few sources of common SVs contain almost exclusively copy-number variants (large indels) and almost none of the more complex SVs such as inversions and translocations. This makes filtering of tumor variants when there is no MN sample very challenging. The Database of Genomic Variants (DGV) [13] is currently the largest SV database, with approximately 200,000 variants. However, >99% of these variants are copy-number changes (gains or losses), and the database contains only a limited number of the more complex structural variant types such as inversions and (balanced) translocations, while these are thought to be common in the normal population [14, 0, 1, 15]. We additionally filtered using BreakSeq database [16], BreakDB [17], and the 1000G SVs [0, 1], which collectively contain another 32,000 structural variants.

We compared the SVs found in the tumor sample to those in the online databases and the VN. We consider two SVs a match if both sides of the event occur within a small distance of the sides of the other SV (200 bp when originating from the same platform, 500 bp when cross-platform).

In the Complete Genomics samples, the VN method removed most of the germline structural variants also removed by the associated normal (~97%) while also removing a further 6%–8% of common variants not corrected for by the associated normal.

For the Illumina samples, the VN method removed fewer SVs than the MN but was still a huge improvement over using only the database filter. The Illumina HCC1187 sample had 132,045 SVs identified in the tumor sample, of which 1464 were of high quality (QUAL ≥ 200). By use of the VN filter, we removed 961 polymorphic variants from this list (Fig. 4.2). Increasing the distance parameter further (to 2000–5000 bp) resulted in the filtering of more SVs than the MN.

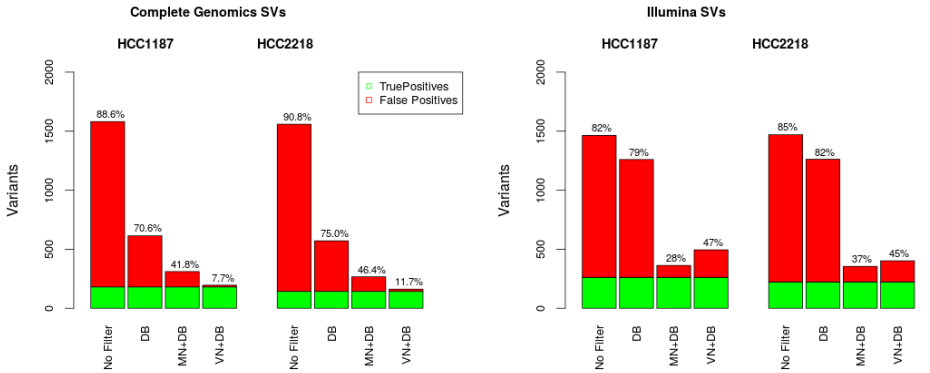


Figure 4.2: Comparison of matched normal (MN) and virtual normal (VN) methods for structural variations (SVs). Correction of high-confidence SVs from Complete Genomics (left) and Illumina (right), using the database filter (DB), MN, and VN. Light gray area indicates the golden set (combination of the three).

	Sample	Detected	Somatic	Description of variants called somatic
Illumina	HCC1187	71 of 98	53	53 of 71 matches survived MN correction; all of these survived DB+ VN correction
Illumina	HCC2218	54 of 64	10	10 of 54 matches survived MN correction; all of these survived DB+ VN correction
CG	HCC1187	91 of 98	91	All survived MN correction; all of these survived DB+ VN correction
CG	HCC2218	55 of 64	55	All survived MN correction; all of these survived DB+ VN correction

Table 4.1: Number of confirmed somatic SVs (as described in COSMIC database) detected in the tumor samples by CG and Illumina, and the number of these variants that are labeled soatic after corrections with our VN method

Experimental validation data for the two public genomes HCC1187 and HCC2218 were obtained from the COSMIC database [18, 19]. We determined the number of these confirmed somatic SVs detected in each sample and determined the number of detected SVs that survived our correction method (Table 4.1). The CG samples had higher sensitivity (detected more of the validated SVs), but for every tumor sample, those variants that were detected by the platform and determined somatic after correction with the MN all survived correction with VN and DB.

SNVs and indels

For the analysis of SNVs and indels for both the MN method and the VN method, we additionally filter variants for their presence in dbSNP [4], the 1000G [0, 1], and the EVS (`http://evs.gs.washington.edu/EVS`) using the ANNOVAR tool [20].

In annotating with the online databases, we require an exact match of position, as well as a match in variant allele between the cancer sample and the variant described in the database. For dbSNP, we used the set of nonflagged variants (flagged variants are those for which SNPs <1% minor allele frequency [MAF; or unknown], mapping only once to reference assembly, or flagged as "clinically associated").

The Illumina FastTrack Cancer Service (http://www.illumina.com/services/whole-genome-sequencing-services/sequencing-service-providers-ign/sequencing-services.ilmn) identified 15,499 somatic SNVs in the HCC1187 sample and 27,823 in HCC2218, after correction with the MN sample. We evaluate performance of our method by correcting the list of all variants found in the tumor sample using our VN set and comparing the remaining variants to those variants determined to be somatic by Illumina's tumor-normal sequencing service.

Two variants are considered a match when they share the same chromosomal position, as well as the same variant allele. Because variants can often be described in various different yet equivalent ways, we used a more advanced correction method for those comparisons involving CG data (the TestVariants tool from CG's tool suite). Detecting these equivalencies is very important for variant comparisons and is discussed in more detail in a later section.

Variants remaining after application of all three filters (MN, database filter, and VN) represent the golden set of true somatic variants (11,409 SNVs for HCC1187 and 20,560 for HCC2218). We determined the number of false-positive somatic variants identified by several filter combinations (Fig. 4.3; Supplemental Data S2). High-confidence variants determined by the VN method are those variants not present in any of the VN samples, and the position may not be no-called in more than ~50% of the samples. This is similar to the MN correction, where typically only those variants are reported that were called reference in the normal sample; tumor variants at positions that are no-called in the normal are usually not reported as (high-confidence) somatic variants.

The VN approach has similar performance to a MN for SNVs and indels and even removes more variants than the associated normal in some cases (Fig. 4.3). However, be aware that this does not mean it removes all of the same variants as the normal but rather it removes an equally large, but different set of variants. There are always highly personal germline variants that can only be removed by the MN, but similarly, there are also polymorphisms that are only removed by the VN and not the single MN sample.

Significant improvement is made over the situation where no MN is available and only aggressive filtering with public databases is used. The advantage of using a VN method rather than relying

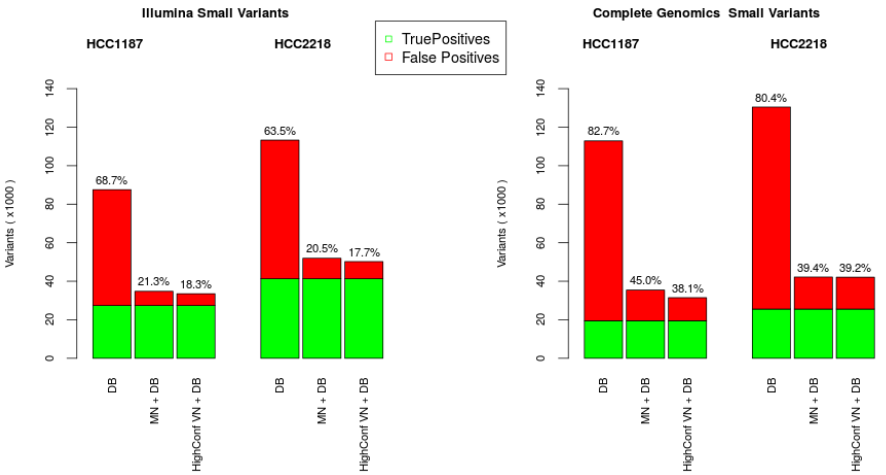


Figure 4.3: Number of false-positive SNVs and indels identified per filtering method for Illumina (left) and Complete Genomics (right). True positives (green) are those variants remaining after application of all filters (for VN, we did not use the high-confidence criterion to determine the set of true positives). DB denotes an aggressive database filter. HighConf VN+DB denotes the list of high-confidence somatic variants as determined by the VN and database filters. MN + DB denotes the list of high-confidence somatic variants after correction with a MN combined with the database filter.

solely on databases is greatest for indels, which are less abundant in the public databases than SNVs and are more difficult to annotate using a purely position-based method because they can often be called in various different but equivalent ways.

To ascertain the quality of the somatic variants identified by our method, we determined several metrics such as Ti/Tv ratio (Supplemental Data S3). We see that the Ti/Tv ratio decreases as more filtering is performed, which is expected for the tumor-specific mutations, as these are more random in nature. Breast cancer specifically has been shown to favor transversion variants [21]. Mutational spectra of the somatic variants determined by our method were also investigated (Supplemental Data S3) and are consistent with the literature [22] in terms of mutation patterns.

Validation data were obtained from COSMIC for both samples. We used this list to determine the number of these validated variants that were detected in the tumor samples by each platform and to determine how many survived correction by our VN method (Table 4.2). Over 94% of the validated variants were detected in each tumor data set, and of the detected variants, only one variant in one of the samples was filtered out only by the VN, indicating a possible false-negative of our method. One confirmed somatic variant in the HCC1187 sample (both Illumina and CG) was present in the associated normals, the public databases, and the VN (17 samples), indicating a possible false

	Sample	Detected	Somatic	Description of variants called somatic
Illumina	HCC1187	82 of 86	75	Five filtered by dbSNP; one by VN only(1x); 1 N + DB + VN(17x)
Illumina	HCC2218	178 of 182	172	Five filtred by dbSNP; one in DB (dbSNP + 1kG + EVS) +VN(19x)
CG	HCC1187	82 of 86	76	Five filtered by dbSNP; one N + DB + VN(17x);
CG	HCC2218	173 of 182	167	Five filtered by dbSNP; one in DB (dbSNP + 1kG + EVS) + VN(9x)

Table 4.2: Number of confirmed somatic variants (as described in COSMIC database) detected in the tumor samples by CG and Illumina, and the number of these variants that are labeled soatic after corrections with our VN method

positive in the COSMIC database. One of the confirmed variants in COSMIC for the HCC2218 sample appeared in our VN nine times, as well as in the dbSNP, the EVS, and 1000G, for both Illumina and CG, indicating another possible false positive in COSMIC. For each of the samples, five confirmed somatic variants were present in the dbSNP (NonFlagged) database. These variants were not present in the VN, the 1000G data, or the EVS and thus are likely disease-related variants that should have been flagged in dbSNP but were not.

Size of VN

Structural variations

We determined the number of VN samples required for filtering of common germline SVs for each sample (Fig. 4.4). For the Complete Genomics samples, after using about 50 VN samples, the same numbers of SVs are filtered out as when using the MN. A plateau is reached after about 120 VN samples, and adding additional normal samples filters out only a small number of additional SVs. Correction with the VN did not remove as many variants as the MN for the Illumina sample, though it still provides significant improvement over filtering with databases alone. A plateau is also reached for the Illumina samples at around 300 VN samples. Increasing the distance threshold for when to consider two SVs a match to 2000 - 5000 bp did result in correction of the same number of SVs as the MN but is likely less accurate.

SNVs and indels

We investigated how many VN samples are necessary for adequate filtering (Fig. 4.5). We are able to attach a confidence measure to the remaining somatic variants by determining the number of VN samples that are no-called at the variants' locus (e.g., a variant at a position that was fully called reference in all 931 normals is more likely to be a true somatic variant than a variant that was also not detected in any of the VN samples but no-called in all normals).

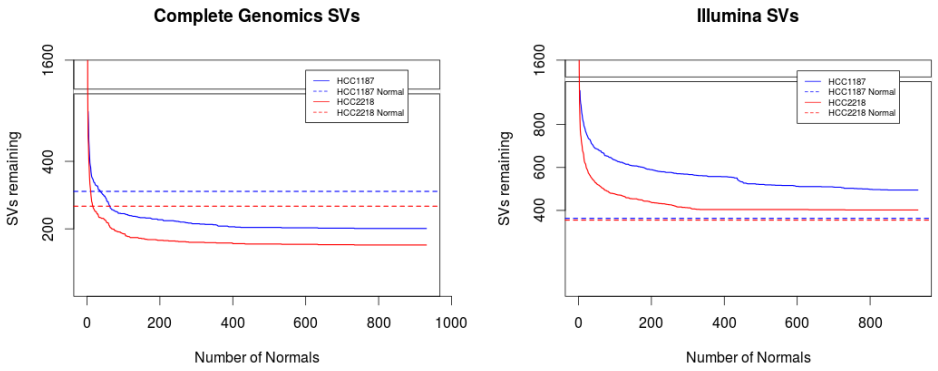


Figure 4.4: Number of structural variants filtered out after each additional VN sample for Complete Genomics (left) and Illumina samples (right). Blue denotes the HCC1187 sample; red, CG HCC2218. Dashed lines indicate the level reached by correction with the associated normal.

The SNV and indel analysis results separated by variant type are presented in Supplemental Data S8. The advantage of the VN method over a database correction was greatest for indels, likely due the fact that these variants can more often be described in various different ways and at slightly different positions and therefore benefit more from the enhanced correction algorithm used in our method than the SNVs. For both SNVs and indels, the number of variants removed is comparable to the number removed by the MN and is significantly more than database corrections alone.

Analysis of SNVs revealed that for the CG samples, about 200 samples are needed to obtain the same performance as using the online databases alone, and any additional genomes added beyond that point will improve the filtering of common variants even further. The Illumina samples required a greater number of CG normals, about 300. For indel variants, the number of normal needed to surpass performance of the online databases alone is fewer than 100 for both platforms. Indel variants can more often be described in different but equivalent forms than SNVs, which means they benefit most from the enhanced comparison method used in our VN approach. The performance is better for Complete Genomics samples than for Illumina samples, but for both platforms, the performance is significantly improved by not relying solely on public mutation databases.

Prostate Cancer Samples

ince HCC1187 and HCC2218 samples are both cell lines, we analyzed two prostate cancer patient samples, G110 and G316, to demonstrate that our method also works for patient data. The results are described in more detail in the Supplemental Data S7. These two samples were sequenced by Complete Genomics and use genome build hg18. Out of our 931 VN samples, 85 were also available

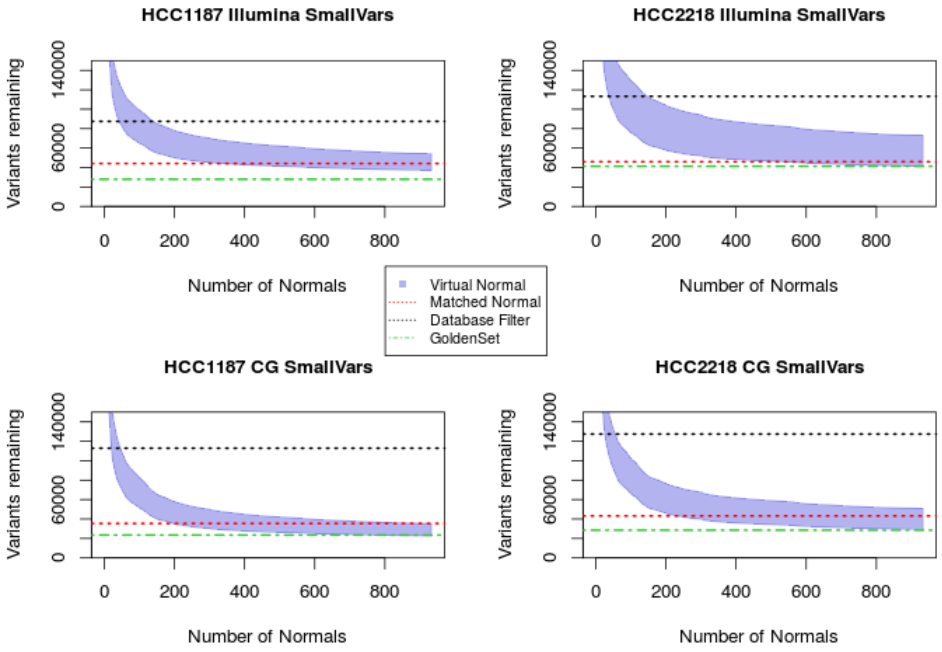


Figure 4.5: Number of SNVs and indels removed after filtering with each additional VN sample. Black dashed line indicates the number of variants labeled as somatic when using only a database filter; red dashed line, the number of variants after correction with MN and the public databases; and the green dashed line, the golden set variants, those remaining after application of all correction methods. The shaded area indicates the number of variants remaining after VN filtering, ranging from all variants (upper bound) to highest-confidence somatic variants (lower bound).

on hg18, so for this analysis we used a smaller VN. Despite this reduced number of normals, our method could still correct as many variants as the MN, requiring around 60–100 normals genomes to do so (20–40 for SVs).

Influence of ethnicity

The influence of ethnicity on the correction power was checked for 54 VN genomes from the Complete Genomics' diversity panel. This panel contains individuals from five different populations across the world. We found that while there was a clear difference between genomes from different races, this difference was <10%, and the set as a whole is capable of correcting as efficiently as a MN regardless of the background of the individual (Supplemental Data S6).

Improved correction method

When comparing the variants found in a (tumor) sample to lists of known polymorphisms, the method most often used is to compare start and end coordinates of the two variants, as well as the observed sequence, and if these three values are identical, the variants are considered equal. However, this approach may be too naïve as there are often various different ways of describing the same variant, depending on the surrounding (reference) sequence and nearby variants.

Consider the following very simple example: Given a reference sequence of CAG and a variant sequence of CAAG, do we describe the variant as an A having been inserted after the C or before the G? Both descriptions are equally valid, but the position of the inserted nucleotide will differ by one. There is no real community consensus on how to resolve these kinds of canonicalization issues, though left alignment is by far the most commonly used. However, the HGVS recommendations urge *"for all descriptions [to use] the most 3′ position possible"*, which would imply right alignment of variants on forward strand genes (`http://www.hgvs.org/mutnomen/recs-DNA.html`).

Another example, and one of the most frequently observed types of annotation difficulties in our data sets, are so-called block substitutions. Whenever two SNVs occur within a 3-bp window and are observed on the same reads, Complete Genomics will call a single 2- or 3-bp substitution, while Illumina will simply call two SNVs. Complete Genomics chooses this approach in order to retain the knowledge that these two variants were encountered on the same allele and possibly within the same codon, which is important for the determination of the impact of the variant on the protein. While the observed sequence for both platforms was the same, the descriptions differ, and naïve comparison methods will not be able to detect the equivalence of these variants.

We identified four main classes of equivalent, but differently described SNVs and indels in our data sets:

1. **Multivariants.** A series of nearby SNVs and indels may also be described as a single, larger substitution in order to retain the knowledge that they occurred on the same allele. The block substitutions are an example of this class (Fig. 4.6A).
2. **Subvariants.** A variant is present in both samples, but in one of the samples, it was adjacent to another variant so in that sample the variant was part of a larger variant (Fig. 4.6B).
3. **Canonicalization.** The same variant sequence is observed in both samples, but it can be described at a different position and possibly with a different observed variant sequence (Fig. 4.6C).

4. **Annotation issues.** Different variant callers and different file formats will have differences in the way they describe variants; for example, when there are multiple variants at the same locus, the descriptions in the VCF format will be different than had the variants occurred alone, turning SNVs into multinucleotide variants and turning simple indels into complicated descriptors (Fig. 4.6D). For example, in the Illumina VCF files, variants of the following pattern are frequently observed: TCA → TA,TC. This indicates that on one allele, the C nucleotide was deleted; on another, the A nucleotide. However, had only the deletion of A occurred, the variant would have been described as CA → C, and thus, the position of the variant would also have been shifted by one. Similarly, had only the deletion of C occurred, the description of the variant would have been TC → T, with an unchanged position field. Many tools converting VCF to a one-line-per-variant format simply split on the comma (TCA → TA and TCA → TC). While it is usually not difficult to reduce these variants to their canonical forms, many tools do not handle this issue correctly, and databases may not always ensure that only canonical forms are entered.

When comparing variants originating from the exact same sequencing and processing pipelines, these issues are minimal, but when comparing variants from different sources, they become more pronounced and must be dealt with in order to maximize the utility of variant databases. We encountered these problems many times when doing comparisons to COSMIC variants and describe several examples in more detail in Supplemental Data S4.

The comparison algorithm we use for our VN correction (CGATools) is capable of detecting most of these equivalences between SNVs and indels and therefore reduces the number of false-positive somatic variants identified.

The description of SVs is even less standardized, making comparisons of variants originating from different sources even more challenging. The differences in calling conventions for SVs are discussed in Supplemental Data S4.

Tumor analysis in Galaxy

Galaxy is a free and open-source web-based analysis platform for data intensive biomedical research [9, 10, 11]. Our VN filtering method is available as a tool for the Galaxy platform as part of our TAG tool suite. The tool can be installed to a local Galaxy instance via the DTL (Dutch Techcenter for Life Sciences) tool shed (`http://toolshed.dtls.nl`). Additional normal samples can easily be added to the VN set in this tool. Further installation and usage instructions can be found within the tool's tool shed repository. The tools have been installed on our demo galaxy

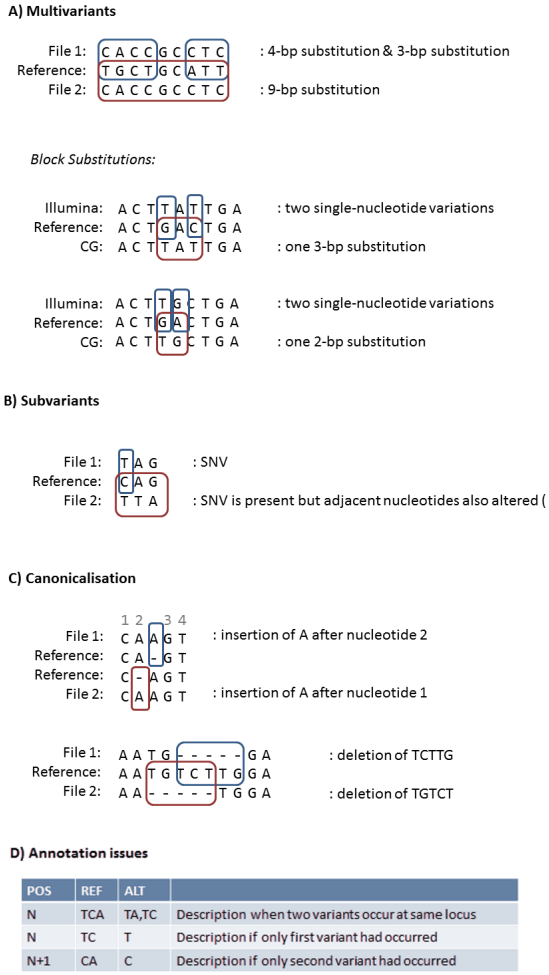


Figure 4.6: Examples of equivalent, but differently annotated variants. (A) When several nearby bases are changed, this can be described as one large substitution or as several smaller ones, even though the resulting sequence is the same. (B) Variant was present, but changes were described as part of a larger variant. (C) Canonicalization issues: Variants can often be described at various different positions, and variants originating from different sources may use different conventions, which must be taken into account during comparisons. (D) In the VCF format, overlapping variants can result in a different description of variants than had they occurred in isolation, a subtlety not always dealt with correctly in comparison algorithms.

example (http://galaxy-demo.ctmm-trait.nl/u/saskia-hiltemann/p/virtual-normal-analysis) [12]; however, due to limited resources, we have had to impose disk and job quotas and recommend installing the tool onto a local (production) Galaxy server for optimal performance. Information about installing and maintaining a Galaxy server is available from the Galaxy wiki (http://galaxyproject.org).

Discussion

We have developed a method for the filtering of tumor variants in the absence of a MN sample. To this end, we have constructed a VN consisting of a set of 931 whole genomes from healthy, unrelated individuals (433 sequenced by Complete Genomics, 498 by Illumina). We evaluated our method on four tumor-normal pairs of two different cancer types, from two different sequencing platforms (CG and Illumina), for both SVs and SNVs and indels. We found that such a VN can correct as many variants as a MN (and in many cases even more), allowing it to possibly serve as a substitute for a MN sample in a research context or provide a valuable addition to the MN in a more clinical setting where highest accuracy is required. It offers a huge improvement over the use of public databases alone, for example, in situations where no normal tissue is available.

Germline variations detected in these tumors after correction with associated normal are in the range of 80%–85% for SVs and 90%–96% for SNVs and small indels and substitutions. Our VN method is able to filter out most of these germline variants (96%–99%) and removes a large number of additional common variants not detected by the associated normal sample. The consensus set of variants—those remaining after correction with associated normal, VN, and public mutation databases—represents the set of true somatic variants. Our method identifies ~10%–30% false-positive SVs and 20%–30% false-positive SNVs and indels, while the tumor-normal method has ~20%–50% and 40%–45% false positives for SVs and SNVs and indels, respectively. This suggests that a VN could act as a substitute for an associated normal when the latter is unavailable, and even outperforms the standard MN correction in terms of false-positive rate in some samples.

The reason for the observation that correction using a VN collection can outperform the MN could be that the sequencing of a single MN sample will not call all germline variants. At the moment with an approximately 100× coverage single-sample analysis by Complete Genomics, ~2.5% of bases (~70,000,000 bases) are not called. If these noncalls are random, a collection of samples will always outperform the correction using a single control. Therefore, we conclude that at the current coverage of 100× or less, a single normal matched control for correction of germline variants should be supplemented with (or even replaced by) a series of at least 200 control genomes in order to deliver optimal results.

We also investigated the optimal size of the VN set and determined that approximately 200–400 genomes are required in order to correct at least as efficiently as the MN sample for SNVs (fewer were required for CG-sequenced samples, as our VN also consisted of CG-sequenced samples). For SVs, 10–40 genomes were required for the CG samples, and for the Illumina samples, our 433

VN samples corrected fewer variants than the associated normal, largely owing to the fact that the description of breakpoints differs so greatly between the two platforms; were we to construct a VN of Illumina-sequenced samples, results would likely vastly improve.

Furthermore, we argue that using a purely position-based annotation incurs additional false positives, should be replaced by an algorithm capable of detecting the equivalence of variants called in different forms, and has knowledge of the reference genome and other nearby variants. When comparing variants obtained from the same sequencing platform and called using the same algorithm, a purely position-based method could suffice, but public databases often contain variants from various sources and are called using various different (versions of) algorithms; in order to utilize the full power of these databases, algorithms need to be able to detect equivalence of variants called in different forms and/or at different locations.

Some caveats to this approach exist: a VN approach cannot correct for those germline variants that are highly personal. Therefore, additional care must be taken when submitting variants to public databases to ensure the anonymity of the patient. In addition, the possibility exists that by foregoing the sequencing of a normal sample, rare germline variants may be mistakenly labeled as somatic, which may lead to rare heritable mutations being overlooked. Additional validation of the somatic status of the variants may be desirable in these cases, either by work in the laboratory or by considering larger groups of patients.

We are able to attach a confidence measure to the somatic variants as determined by our VN method by considering the number of VN samples that are no-called at the variants' loci. A somatic variant at a position that was fully called reference in all 931 normals is more likely to be a true somatic variant than a variant that was also not detected in any of the VN samples but was no-called in all normal samples (i.e., evidence of absence vs. absence of evidence). Currently, this could only be done for the VN samples sequenced by CG, as the Illumina data does not provide the necessary information about no-called and half-called loci.

Our VN correction with the 931 samples can be run on any VCF file or list of variants, regardless of the sequencing platform of origin, and is available as a Galaxy tool from the DTLS tool shed (`http://toolshed.dtls.nl/repos/saskia-hiltemann/virtual_normal_preprocessing`) and is installed on our public demonstration Galaxy server (`http://galaxy-demo.ctmm-trait.nl`).

Methods

Illumina vs. Complete Genomics

For this study we analyzed two breast cancer cell line samples, HCC1187 and HCC2218, which have been whole-genome-sequenced by both Complete Genomics and Illumina and are publicly available for download (ftp://ftp2.completegenomics.com/Cancer_pairs/ and https://basespace.illumina.com/datacentral).

Comparisons between the Illumina and Complete Genomics platforms have been previously described [23]. We compared the variants identified by each of the platforms for our samples and found that >96% of the SNVs and >70% of the indels identified by Complete Genomics were also present in the Illumina samples (Supplemental Data S1).

The description of SVs differs greatly between the two platforms, making comparison a challenging task. Our algorithm found an overlap of ~40%–45% (Supplemental Table S1.3). The calling differences and comparison pitfalls are discussed in a later section.

Complete Genomics samples

The HCC1187 breast cancer (primary ductal carcinoma) sample was TNM stage IIA, grade 3. For HCC1187 BL, the normal sample was derived from peripheral blood and immortalized with EBV transformation. ATCC numbers were as follows: tumor, CRL-2322; normal, CRL-2323. CG Software version used was 2.0.2.15.

The HCC2218 breast cancer (primary ductal carcinoma) sample was TNM stage IIIA, grade 3. For NA12880, the normal sample was derived from peripheral blood and immortalized with EBV transformation. ATCC numbers were as follows: tumor, CRL-2343; normal, CRL-2363. CG Software version used was 2.0.2.15.

Samples have been sequenced to an average genome-wide coverage of 123× for three of the samples and 92× for the NA12880 sample. The HCC1187 and HCC2218 samples were downloaded from Complete Genomics (ftp://ftp2.completegenomics.com/Cancer_pairs/).

The prostate cancer sample G110 is derived from a radical prostatectomy. The tumor section from which DNA was isolated had a Gleason score of 3 + 3 and contained 80% epithelial tissue, of which 90% was cancer. The MN DNA was isolated from peripheral blood. The G110 and MN samples were sequenced by Complete Genomics to an average genome-wide coverage of 94× and 109×, respectively. The software version used was 2.0.2.24.

The prostate cancer sample G316 is derived from a transurethral resection of the prostate (TURP). The tumor section from which DNA was isolated had a Gleason score of 4 + 3 and contained only epithelial tissue, of which 100% was cancer. The MN DNA was isolated from peripheral blood. The G316 and MN samples were sequenced by Complete Genomics to an average genome-wide coverage of 112× and 113×, respectively. The software version used was 2.0.2.24. VN samples

The 433 normal samples were sequenced by Complete Genomics in the context of the 1000G and are accessible for download from the EBI and NCBI ftp servers at `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/` or `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/`.

For the hg18 prostate cancer samples, our VN set contained 66 normals, consisting of (1) the Complete Genomics diversity panel (46 genomes), (2) the four unrelated individuals from the CG pedigree, (3) the parents in the two CG trios (YRI and PUR), and (4) 12 in-house samples of healthy, unrelated individuals. This amounts to a total of 66 genomes. For the SNV and indel analysis, an additional 19 in-house samples were available, bringing the total up to 85. The Complete Genomics public samples may be downloaded from `ftp://ftp2.completegenomics.com/`. A list of all variants found in our in-house VN samples is available as a (public) shared data set from our demonstration Galaxy server (`http://galaxy-demo.ctmm-trait.nl`) and in the Supplemental Materials.

For the Illumina normals, sample-level data were obtained from the GoNL Consortium (`http://www.nlgenome.nl`) to perform this analysis. Our tool will only output summary counts for these data as the individual-level data are restricted. Comparison to COSMIC validated variants

The validated variants obtained from COSMIC used hg19 coordinates, while our in-house samples were sequenced on hg18. The public CG samples were sequenced on both hg18 and hg19, so for this comparison, we used a VN consisting of just the 54 public hg19 genomes because a lift-over of genomic coordinates is suboptimal.

SV analysis

We used CGATools JunctionDiff version 1.6 with default parameter settings for both the tumor-normal filtering and the tumor-VN filtering of the Complete Genomics samples. This means we considered two junctions to be the same when both the left sides and the right sides of the two junctions are on the same strand and fall within 200 bp of each other. For comparisons involving Illumina SVs, we created a custom script labelling two SVs as similar if they fall within a short distance of each other (for both sides of the event). We used a distance of 500 bp if the events

came from the same platform, 1000 if they came from different technologies.

The CGATools source code and binaries are freely available for download at http://cgatools.sourceforge.net.

HCC1187 validation data in COSMIC are available at http://cancer.sanger.ac.uk/cosmic/sample/overview?id=749711. Data describe breast tissue, and the carcinoma is ductal. The number of genes examined is 4675; simple mutations, 29; gene fusions, 12; and structural variants, 94.

HCC2218 validation data in COSMIC are available at http://cancer.sanger.ac.uk/cosmic/sample/overview?id=749716. Data describe breast tissue, and the carcinoma is ductal. The number of genes examined is 4670; simple mutations, 76; gene fusions, 0; and structural variants, 62.

SNV and indel analysis

We used the Complete Genomics CGATool ListVariants and TestVariants (version 1.6) for our SNV and indel analysis when comparisons involved CG data. When comparing Illumina tumors to Illumina VNs, we used vcftools (http://vcftools.sourceforge.net) and vcflib (http://github.com/ekg/vcflib), which compares positions as well as observed variant sequence when determining a match.

We used ANNOVAR (release date 2013 Feb 11) for annotation with the public variant databases. In annotating with the online databases, we require an exact match of position, as well as a match in variant allele between the cancer sample and the variant described in the database. For dbSNP, we used the set of nonflagged variants (flagged variants are those for which SNPs <1% MAF (or unknown), mapping only once to reference assembly, or flagged as "clinically associated").

HCC1187 validation data are available at http://cancer.sanger.ac.uk/cosmic/sample/overview?id=1235080. Data describe breast tissue, and the carcinoma is ductal. The number of genes examined is 12,196; simple mutations, 55; gene fusions, 0; and structural variants, 0.

HCC2218 validation data are available at http://cancer.sanger.ac.uk/cosmic/sample/overview?id=1235085. Data describe breast tissue, and the carcinoma is ductal. The number of genes examined is 12,196; simple mutations, 107; gene fusions, 0; and structural variants, 0.

dbSNP annotations are from the Database of Single Nucleotide Polymorphisms (dbSNP), National Center for Biotechnology Information, National Library of Medicine (dbSNP Build ID: all builds up to 38), and are available at http://www.ncbi.nlm.nih.gov/SNP/.

Data access

WGS variation data from this study are available at the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena), accession number PRJEB9673. The list of variants present in our 31 in-house normal samples can be found in the Supplemental Material.

Acknowledgments

We thank Rick Tearle and Steve Lincoln from Complete Genomics, whose valuable discussions on Complete Genomics analysis methods supported our study. This study was performed within the framework of the Center for Translational Molecular Medicine (CTMM), TraIT project (grant 05T-401). This study makes use of data generated by the Genome of the Netherlands Project. A full list of the investigators is available from www.nlgenome.nl. Funding for the project was provided by the Netherlands Organization for Scientific Research under award no. 184021007, dated July 9, 2009, and made available as a Rainbow Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL). The sequencing was carried out in collaboration with the Beijing Institute for Genomics (BGI). This work was sponsored by the BiG Grid project for the use of the computing and storage facilities, with financial support from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organization for Scientific Research, NWO).

Bibliography

[0] . G. P. Consortium *et al.*, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[1] . G. P. Consortium *et al.*, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, p. 56, 2012.

[2] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat, "Sensitive and accurate detection of copy number variants using read depth of coverage," *Genome research*, vol. 19, no. 9, pp. 1586–1592, 2009.

[3] A. Kumar, T. A. White, A. P. MacKenzie, N. Clegg, C. Lee, R. F. Dumpit, I. Coleman, S. B. Ng, S. J. Salipante, M. J. Rieder, *et al.*, "Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers," *Proceedings of the National Academy of Sciences*, vol. 108, no. 41, pp. 17087–17092, 2011.

[4] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbsnp: the ncbi database of genetic variation," *Nucleic acids research*, vol. 29, no. 1, pp. 308–311, 2001.

[5] R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, *et al.*, "Human genome sequencing using unchained base reads on self-assembling dna nanoarrays," *Science*, vol. 327, no. 5961, pp. 78–81, 2010.

[6] D. I. Boomsma, C. Wijmenga, E. P. Slagboom, M. A. Swertz, L. C. Karssen, A. Abdellaoui, K. Ye, V. Guryev, M. Vermaat, F. Van Dijk, *et al.*, "The genome of the netherlands: design, and project goals," *European Journal of Human Genetics*, vol. 22, no. 2, p. 221, 2014.

[7] L. C. Francioli, A. Menelaou, S. L. Pulit, F. Van Dijk, P. F. Palamara, C. C. Elbers, P. B. Neerincx, K. Ye, V. Guryev, W. P. Kloosterman, *et al.*, "Whole-genome sequence variation, population structure and demographic history of the dutch population," *Nature genetics*, vol. 46, no. 8, p. 818, 2014.

[8] V. Boeva, A. Zinovyev, K. Bleakley, J.-P. Vert, I. Janoueix-Lerosey, O. Delattre, and E. Barillot, "Control-free calling of copy number alterations in deep-sequencing data using gc-content normalization," *Bioinformatics*, vol. 27, no. 2, pp. 268–269, 2010.

[9] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko, "Galaxy: a platform for interactive large-scale genome analysis," *Genome Res*, vol. 15, no. 10, pp. 1451–1455, 2005.

[10] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor, "Galaxy: A web-based genome analysis tool for experimentalists," *Curr Protoc in Mol Biol*, pp. 19.10.1 – –19.10.21, 2010.

[11] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome biology*, vol. 11, no. 8, p. R86, 2010.

[12] S. Hiltemann, H. Mei, M. de Hollander, I. Palli, P. van der Spek, G. Jenster, and A. Stubbs, "Cgtag: complete genomics toolkit and annotation in a cloud-based galaxy," *GigaScience*, vol. 3, no. 1, p. 1, 2014.

[13] J. R. MacDonald, R. Ziman, R. K. Yuen, L. Feuk, and S. W. Scherer, "The database of genomic variants: a curated collection of structural variation in the human genome," *Nucleic acids research*, vol. 42, no. D1, pp. D986–D992, 2013.

[14] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics*, vol. 7, no. 2, p. 85, 2006.

[15] R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, *et al.*, "Mapping copy number variation by population-scale genome sequencing," *Nature*, vol. 470, no. 7332, p. 59, 2011.

[16] H. Y. Lam, X. J. Mu, A. M. Stütz, A. Tanzer, P. D. Cayting, M. Snyder, P. M. Kim, J. O. Korbel, and M. B. Gerstein, "Nucleotide-resolution analysis of structural variants using breakseq and a breakpoint library," *Nature biotechnology*, vol. 28, no. 1, p. 47, 2010.

[17] J. O. Korbel, A. Abyzov, X. J. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder, and M. B. Gerstein, "Pemer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data," *Genome biology*, vol. 10, no. 2, p. R23, 2009.

[18] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, *et al.*, "Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D945–D950, 2010.

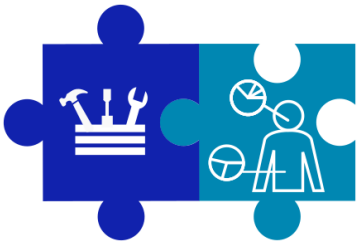
[19] N. Bindal, S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, C. Y. Kok, M. Jia, S. Bamford, C. Cole, S. Ward, *et al.*, "Cosmic: the catalogue of somatic mutations in cancer," *Genome biology*, vol. 12, no. 1, p. P3, 2011.

[20] K. Wang, M. Li, and H. Hakonarson, "Annovar: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic acids research*, vol. 38, no. 16, pp. e164–e164, 2010.

[21] S. Liu, W. Liu, J. L. Jakubczak, G. L. Erexson, K. R. Tindall, R. Chan, W. J. Muller, S. Adhya, S. Garges, and G. Merlino, "Genetic instability favoring transversions associated with erbb2-induced mammary tumorigenesis," *Proceedings of the National Academy of Sciences*, vol. 99, no. 6, pp. 3770–3775, 2002.

[22] A. F. Rubin and P. Green, "Mutation patterns in cancer genomes," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21766–21770, 2009.

[23] H. Y. Lam, M. J. Clark, R. Chen, R. Chen, G. Natsoulis, M. O'huallachain, F. E. Dewey, L. Habegger, E. A. Ashley, M. B. Gerstein, *et al.*, "Performance comparison of whole-genome sequencing platforms," *Nature biotechnology*, vol. 30, no. 1, p. 78, 2012.



"The more clearly we can focus our attention on the wonders and realities of the universe about us, the less taste we shall have for destruction."

Rachel Carson

5

Microbiota Profiling

With sequencing costs steadily decreasing, 16S sequencing has become a viable alternative to culture-based methods in routine clinical diagnostics. For any clinical application, extensive validation is required. To pilot the utility of 16S profiling in a clinical setting, we first integrated the full mothur toolsuite into Galaxy (GmT), and subsequently used this tool suite to create a set of workflows for clinical experimental designs in collaboration with Streeklab Haarlem (MYcrobiota).

This chapter contains the following sub-chapters:

5.0 **Galaxy mothur Toolset (GmT): a user-friendly application for 16S rRNA gene sequencing analysis using mothur** Our aim was to provide a data analysis method that could be used by clinicians directly. In diagnostics, there is often a high throughput of analysis, which must be run in a timely fashion, so if we enable clinicians to run this analysis themselves, without needing to wait for a bioinformatician, we can potentially speed up the analysis process significantly. The first step towards this goal was to integrate the required tools into Galaxy. We identified the mothur tool suite as the optimal solution for your use case, however this package contained over 125 tools. I wrapped all of these into Galaxy and created workflows and tutorials for the recommended Standard Operating Procedure (SOP).

5.1 **MYcrobiota: Development and evaluation of a culture-free microbiota profiling platform for clinical diagnostics.** After the required tools were made available in Galaxy (previous sub-chapter), we started working on adapting the SOP to fit the needs of the Streeklab Haarlem. This involved an combination of bioinformatics analysis (by me) and experimental validation (Stephan Boers). The novel experimental approach employed by the Streeklab Haarlem (micelle PCR), as well as experimental setup (sequencing in triplicate, negative extraction control, mock sample sequencing) required corresponding adaptations to the standard workflow. Furthermore, we created a customized analysis report to display to clinicians based on the iReport tool. And of course extensive validation was needed before the MYcrobiota platform could be considered for adoption into routine diagnostics. Since publication of this work, the MYcrobiota has continued to be actively used by the Streeklab Haarlem.

Galaxy mothur Toolset (GmT): a user-friendly application for 16S rRNA gene sequencing analysis using mothur.

Saskia Hiltemann[1,*], Stefan Boers[2,*], Peter van der Spek [1], Ruud Jansen[3], John Hays[2], Andrew Stubbs[1]

1. Department of Bioinformatics, Erasmus Medical Center, Rotterdam, The Netherlands.
2. Department of Medical Microbiology and Infectious Diseases, Erasmus Medical Centre, Rotterdam, The Netherlands
3. Department of Molecular Biology, Regional Laboratory of Public Health Kennemerland, Haarlem, The Netherlands.

Published in: *GigaScience*, 2019 Feb 1;8(2):giy166

DOI: https://doi.org/10.1093/gigascience/giy166

*: Saskia D. Hiltemann and Stefan A. Boers contributed equally to this work.

Abstract

Background The determination of microbial communities using the mothur tool suite (https://www.mothur.org) is well established. However, mothur requires bioinformatics-based proficiency in order to perform calculations via the command line. Galaxy is a project dedicated to providing a user-friendly web interface for such command line tools (https://galaxyproject.org/).

Results: We have integrated the full set of 125+ mothur tools into Galaxy as the Galaxy mothur Toolset (GmT) and provided a set of workflows to perform end-to-end 16S rRNA gene analyses and integrate with third-party visualization and reporting tools. We demonstrate the utility of GmT by analysing the mothur MiSeq standard operating procedure (SOP) data set (https://www.mothur.org/wiki/MiSeq_SOP).

Conclusions: GmT is available from the Galaxy Tool Shed, and a workflow definition file and full Galaxy training manual for the mothur SOP have been created. A Docker image with a fully configured GmT Galaxy is also available

Keywords: Microbial classification; 16S rRNA gene sequence analysis; mothur

Findings

Introduction

rRNA gene profiling analysis can be achieved using an extensive array of sophisticated software including mothur [0], QIIME [1], MG-RAST [2], and many more [3]. Whilst some of these applications have a graphical user interface (GUI) to provide access to these technologies for the research scientist, their use remains complex for non-bioinformaticians. In this respect, the Galaxy project [4] was developed in order to simplify the use of complex command line software tools. Galaxy offers extensive support for both 16S rRNA gene-based and broader metagenomic analyses, with over 100 tools in the metagenomics section of the Galaxy tool shed, including QIIME [1], KRONA [5], PyNAST [6], PICRUSt [7], Kraken [8], MetaPhlAn2 [9], HUMAnN2 [10], PrinSEQ [11], Nonpareil [12], Vegan [13], and many more.

Mothur is an open-source application that was designed as a single piece of software capable of analysing and comparing microbial communities from 16S rRNA gene data derived from next-generation sequencing (NGS). The creators of mothur did not only provide an extensive set of tools, but also a collection of standard operating procedures (SOPs) that detail the recommended analytical protocol for different types of input data.

The latest version of mothur consists of over 125 components, lending it great flexibility, but at the same time, great complexity. To address this challenge, we have integrated the full set of 125+ mothur components into Galaxy that are collectively called the 'Galaxy mothur Toolset (GmT)'. To simplify usage of GmT we provide the full workflow definition files, usage of which shields the end-user from the full complexities of the analysis. By simultaneously providing access to all the individual components present in mothur as separate tools, expert users and bioinformaticians retain the ability to utilize the full flexibility of mothur by creating custom workflows or by modifying or extending our workflows to fit their use-case.

GmT also leverages Galaxy's collections framework to enable easy analysis of large numbers (many thousands) of samples at once. Many mothur components support parallel computing, and the Galaxy tools will utilize the maximum amount of processing power alloted to them by the instance administrator. As part of GmT, datatypes were also contributed to the Galaxy core codebase to facilitate the handling of mothur-specific datatypes within Galaxy. Furthermore, a Galaxy data manager was also created for the automatic installation and configuration of reference datasets utilized by the mothur tool suite. And lastly, a Galaxy Interactive environment (GIE) [14] for Phinch [15] was also developed [16].

GmT includes tools to produce standard file formats, such as the BIOM format [17] to facilitate interoperability with these downstream analysis components. Where no clear file standards exists, GmT provides custom tools for conversion of mothur datatypes to other tools (e.g. the taxonomy-2-krona tool). This allows for integration with third party tools such as PICRUSt for prediction of functional content, or visualisation tools such as Phinch, KRONA, and certain QIIME components. The mothur tools also natively support incorporation of some 3rd party analysis tools, such as UCHIME and ChimeraSlayer for chimera detection or VSEARCH for clustering, which are also available in GmT.

The Galaxy Training Network (GTN) is a network of people and groups that present Galaxy and Galaxy-based training around the world. The GTN has created a central repository [18] for Galaxy training materials. In order to further facilitate the use of GmT to end-users, we have contributed training materials to the GTN that illustrate how to run mothur's MiSeq SOP within Galaxy [19]. This work has also been incorporated in a larger-scale framework to easily and quickly explore microbiota data in a reproducible and transparent environment [20].

Purpose of this work

The work performed and described in this technical note has four objectives. First to provide end-users and bioinformaticians with easy access to all the mothur tools as the GmT. Second is to provide open-access online training material to demonstrate/complete the mothur SOP in Galaxy. Third is to deliver an end-to-end workflow for the mothur SOP in Galaxy that is available for upload to any Galaxy that has the GmT installed. Fourth is to provide a summarization of results in a web report using the iReport Galaxy tool [21]. Our aim is to provide 16S rRNA gene NGS analysis tools and awareness on how to use them in a format that supports FAIR data principles [22].

Worked Example

To illustrate the utility of our toolkit, we present results on example data below. GmT is designed to take short-read 16S rRNA gene NGS data as input and to output a dynamic web report for prokaryotic taxonomical classification using the Galaxy platform. A GmT workflow follows essentially a four-step process:

1. **Data upload**. The Galaxy platform provides the users with standard data upload functionality for single and multi-sample datasets.
2. **Collection Creation**. For multi-sample and/or paired-end datasets a Galaxy collection must be created in the Galaxy interface. Here datasets can also be assigned to groups. Galaxy will make intelligent suggestions for pairings of datasets based on the file names.
3. **16S rRNA gene analysis**. Mothur has been wrapped as a tool suite in Galaxy. Required steps included for a full *end-to-end* 16S rRNA gene sequencing analysis consist of read-pair merging (mothur command: make.contigs), trimming of primer sequences (trim.seqs), additional quality

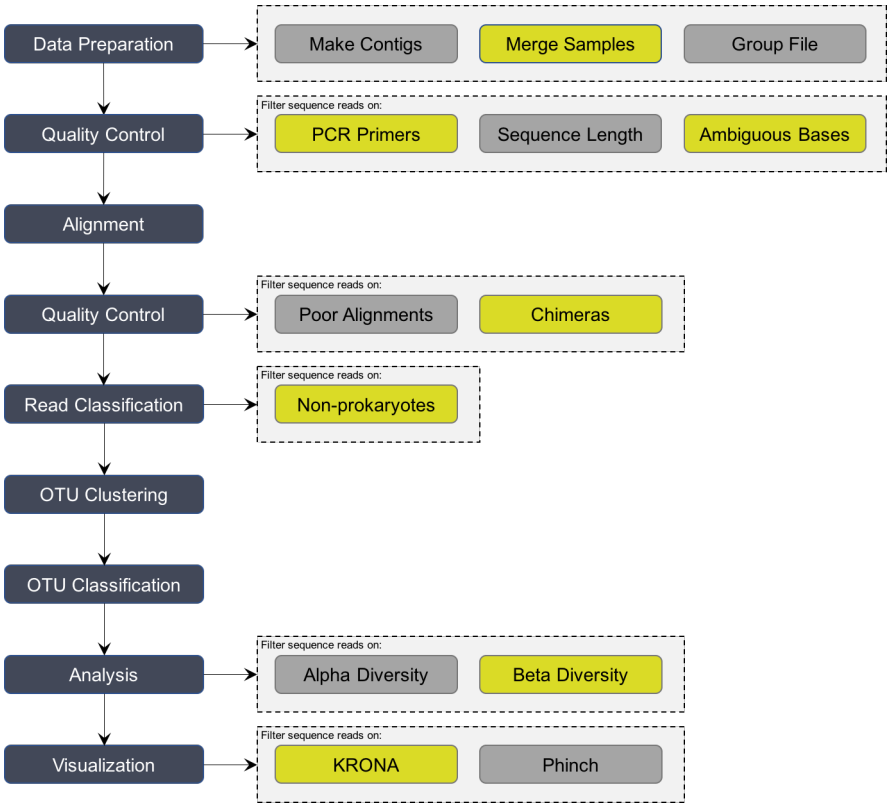
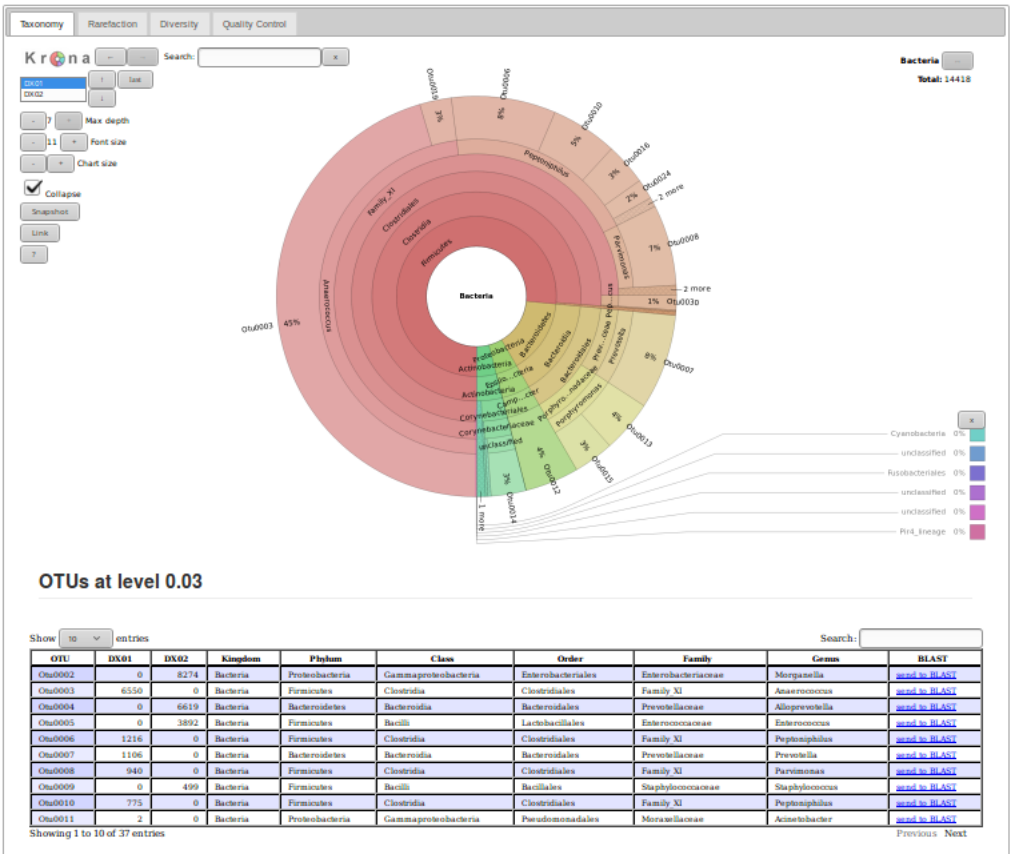


Figure 5.0: Conceptual view of the GmT mothur MiSeq SOP pipeline.

control (screen.seqs), alignment of sequences to a (customized) reference alignment (align.seqs, screen.seqs), removal of chimeric sequences (chimera.uchime), classifying sequences using a Bayesian classifier in combination with a reference database such as SILVA or GreenGenes (classify.seqs), and clustering of sequences into operational taxonomic units (OTUs) at a predefined percentage - usually 97 percent - of similarity (dist.seqs, cluster, and classify.otu) (Figure 5.0).

4. **Experimental Summary and Reporting**. iReport in combination with KRONA is used to deliver an HTML report in Galaxy [5]. The iReport consists of multiple tabs to group results topically (e.g. taxonomy, rarefaction, diversity, quality control) and is highly customizable and easily tailored to an end-user's specific use-case. The entire report may be downloaded from the Galaxy interface to be viewed or shared offline.

   To compare the output from a single experiment or across multiple experiments we utilized Phinch [15], a dynamic web application which uses BIOM-formatted files to explore and analyse biological patterns in 16S rRNA gene NGS datasets.



OTU	DX01	DX02	Kingdom	Phylum	Class	Order	Family	Genus	BLAST
Otu0002	0	8274	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Morganella	send to BLAST
Otu0003	6550	0	Bacteria	Firmicutes	Clostridia	Clostridiales	Family XI	Anaerococcus	send to BLAST
Otu0004	0	6619	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Alloprevotella	send to BLAST
Otu0005	0	3892	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus	send to BLAST
Otu0006	1216	0	Bacteria	Firmicutes	Clostridia	Clostridiales	Family XI	Peptoniphilus	send to BLAST
Otu0007	1106	0	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	send to BLAST
Otu0008	940	0	Bacteria	Firmicutes	Clostridia	Clostridiales	Family XI	Parvimonas	send to BLAST
Otu0009	0	499	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus	send to BLAST
Otu0010	775	0	Bacteria	Firmicutes	Clostridia	Clostridiales	Family XI	Peptoniphilus	send to BLAST
Otu0011	2	0	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	send to BLAST

Figure 5.1: Example iReport. This web report contains the interactive KRONA visualization, the (multi-sample) OTU table, rarefaction plots, diversity calculations, differential abundance analysis, and an extensive overview of the quality control measurements taken during the analysis. iReports are highly customizable and can be easily tailored to fit specific use-cases and end-user needs.

Methods

Handling large datasets

Large-scale analyses have become the norm in the field, both large in disk space as in the number of files, and this can pose a challenge for analysis. For large files, Galaxy offers the option of uploading via FTP rather than web transfer. The introduction of the concept of "collections" in Galaxy has enabled users to analyze datasets consisting of a large number of files (>100K) as easily as they would a single file.

Galaxy mothur Toolset

Many mothur components support parallelization, and our Galaxy wrappers will run these components with the maximum number of CPUs allotted to them by the Galaxy administrator. In order to diagnose potential failures, Galaxy outputs the full standard and error logs, which the users can inspect. Furthermore, we have contributed mothur datatype definitions to the Galaxy core code, meaning that the users will be protected from inputting the wrong datasets and thus reduce the number of errors they will make with the tools. All tools in GmT use only conda dependencies, making their installation in Galaxy a painless experience that requires nothing more than a single press of a button.

The mothur tool wrappers have been submitted to the Intergalactic Utilities Commission (IUC) tool repository [23] and are available from the Galaxy Tool Shed [24]. The IUC is a group of community members dedicated to developing and upholding Galaxy tool development best practices and guidelines, thus by contributing our tools to this repo we ensure that the tools will be well-maintained. A metagenomics Galaxy flavour [25] is available which contains all components presented here. The full mothur suite has also been installed to Galaxy's main server [26].

KRONA visualization

KRONA [5] is a data viewer which provides the ability to interactively explore hierarchical data. A Galaxy KRONA wrapper that works directly on mothur data formats was developed for this project.

Phinch visualization

Galaxy offers integration with Phinch [15] BIOM format viewer in two ways; as a Galaxy interactive environment (GIE) developed in the context of this project [16], and more recently also as an external display application hosted by the Galaxy team.

iReport summarization

To facilitate the evaluation of 16S rRNA gene sequencing analysis results, integration with the iReport [21] tool are also provided. This tool creates a web report to present the analysis results in an organized fashion and provides links to external resources such as BLAST searches (Figure 17).

Availability of source code and requirements

- Project name: Galaxy mothur Toolset (GMT)
- Project home page: `https://github.com/erasmusmc-bioinformatics/galaxy-mothur-toolset`
- Toolshed repository: `https://toolshed.g2.bx.psu.edu/view/iuc/suite_mothur/768c2e48b706`

- Training manual: https://galaxyproject.github.io/training-material
- GmT Docker image: https://quay.io/shiltemann/galaxy-mothur-toolset:16.07
- Galaxy Metagenomics Docker Flavour (Docker): https://quay.io/repository/shiltemann/galaxy-metagenomics, https://github.com/shiltemann/galaxy-metagenomics
- Phinch interactive environment: https://github.com/shiltemann/phinch-galaxy-ie
- Operating system: Unix (Platform independent with Docker)
- License: GNU GPL v3

Availability of supporting data and materials

The data presented here to illustrate our work is the same data used in the training manual, and is available from Zenodo [27]

Declarations

List of abbreviations

- **GIE:** Galaxy interactive environment
- **GmT:** Galaxy mothur Toolset
- **GTN:** Galaxy Training Network
- **GUI:** Graphical user interface
- **SOP:** Standard Operating Procedure

Competing Interests

The authors declare that they have no competing interests.

Funding

This work has received funding from the European Union's Seventh Framework Programme for Health under grant agreement number 602860 (TAILORED-Treatment; www.tailored-treatment.eu).

Author's Contributions

SH developed the Galaxy tool wrappers and Phinch interactive environment. SB validated the analysis pipelines. All authors contributed to the manuscript text and approve its contents.

Acknowledgements

The authors would like to thank Jim Johnson and the many other contributors and reviewers of the mothur tool wrappers, including everybody who contributed to the development of these tools within the context of the Galaxy metagenomics contribution fest organized by the Galaxy community's Intergalactic Utilities Commission (IUC), a group of community members dedicated to developing and upholding Galaxy tool development best practices and guidelines [28]. We would also like to thank the Galaxy Training Network for providing the infrastructure and valuable feedback to share our training materials.

Bibliography

[0] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, *et al.*, "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and environmental microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[1] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, *et al.*, "Qiime allows analysis of high-throughput community sequencing data," *Nature methods*, vol. 7, no. 5, pp. 335–336, 2010.

[2] E. M. Glass, J. Wilkening, A. Wilke, D. Antonopoulos, and F. Meyer, "Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes," *Cold Spring Harbor Protocols*, vol. 2010, no. 1, pp. pdb–prot5368, 2010.

[3] A. Oulas, C. Pavloudi, P. Polymenakou, G. A. Pavlopoulos, N. Papanikolaou, G. Kotoulas, C. Arvanitidis, and I. Iliopoulos, "Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies," *Bioinformatics and Biology insights*, vol. 9, p. 75, 2015.

[4] E. Afgan, D. Baker, M. Van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, *et al.*, "The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W3–W10, 2016.

[5] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Interactive metagenomic visualization in a web browser," *BMC bioinformatics*, vol. 12, no. 1, p. 385, 2011.

[6] J. G. Caporaso, K. Bittinger, F. D. Bushman, T. Z. DeSantis, G. L. Andersen, and R. Knight, "Pynast: a flexible tool for aligning sequences to a template alignment," *Bioinformatics*, vol. 26, no. 2, pp. 266–267, 2009.

[7] M. G. Langille, J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. V. Thurber, R. Knight, *et al.*, "Predictive functional profiling of microbial communities using 16s rrna marker gene sequences," *Nature biotechnology*, vol. 31, no. 9, p. 814, 2013.

[8] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome biology*, vol. 15, no. 3, p. R46, 2014.

[9] D. T. Truong, E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, and N. Segata, "Metaphlan2 for enhanced metagenomic taxonomic profiling," *Nature methods*, vol. 12, no. 10, p. 902, 2015.

[10] S. Abubucker, N. Segata, J. Goll, A. M. Schubert, J. Izard, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, *et al.*, "Metabolic reconstruction for metagenomic data and its application to the human microbiome," *PLoS computational biology*, vol. 8, no. 6, p. e1002358, 2012.

[11] R. Schmieder and R. Edwards, "Quality control and preprocessing of metagenomic datasets," *Bioinformatics*, vol. 27, no. 6, pp. 863–864, 2011.

[12] L. M. Rodriguez-r and K. T. Konstantinidis, "Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets," *Bioinformatics*, vol. 30, no. 5, pp. 629–635, 2013.

[13] P. Dixon, "Vegan, a package of r functions for community ecology," *Journal of Vegetation Science*, vol. 14, no. 6, pp. 927–930, 2003.

[14] H. Rasche, B. Gruening, J. Chilton, and D. Baker, "Galaxy interactive environments–a new way to interact with your data," in *Galaxy Community Conference*, 2015.

[15] H. M. Bik and P. Interactive, "Phinch: An interactive, exploratory data visualization framework for–omic datasets," *bioRxiv*, p. 009944, 2014.

[16] S. Hiltemann, "Phinch galaxy interactive environment." https://github.com/shiltemann/phinch-galaxy-ie.

[17] "The biological observation matrix (biom) format." http://biom-format.org.

[18] "Galaxy training network." https://training.galaxyproject.org.

[19] "Gtn tutorial: 16s microbial analysis with mothur." https://training.galaxyproject.org/training-material/topics/metagenomics/tutorials/mothur-miseq-sop/tutorial.html.

[20] B. Batut, K. Gravouil, C. Defois, S. Hiltemann, J.-F. Brugère, E. Peyretaillade, and P. Peyret, "Asaim: a galaxy-based framework to analyze raw shotgun data from microbiota," *bioRxiv*, p. 183970, 2017.

[21] S. Hiltemann, Y. Hoogstrate, P. Van der Spek, G. Jenster, and A. Stubbs, "ireport: a generalised galaxy solution for integrated experimental reporting," *GigaScience*, vol. 3, no. 1, p. 19, 2014.

[22] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, p. 160018, 2016.

[23] "Tool shed repositories maintained by the intergalactic utilities commission." https://github.com/galaxyproject/tools-iuc.

[24] "Main galaxy toolshed." http://toolshed.g2.bx.psu.edu/.

[25] "Galaxy docker repository for metagenomics (galaxy metagenomics flavour)." https://github.com/shiltemann/galaxy-metagenomics.

[26] "Galaxy main server." https://usegalaxy.org.

[27] "Mothur miseq sop galaxy tutorial data." https://zenodo.org/record/800651.

[28] "Intergalactic utilities commission." https://galaxyproject.org/iuc/.

Development and evaluation of a culture- free microbiota profiling platform (MYcrobiota) for clinical diagnostics

Stefan Boers[2,*], Saskia Hiltemann[1,*], Andrew Stubbs[1], Ruud Jansen[3], John Hays[2]

1. Department of Bioinformatics, Erasmus Medical Center, Rotterdam, The Netherlands.
2. Department of Medical Microbiology and Infectious Diseases, Erasmus Medical Center, Rotterdam, The Netherlands.
3. Department of Molecular Biology, Regional Laboratory of Public Health Kennemerland, Haarlem, The Netherlands.

Published in: European Journal of Clinical Microbiology & Infectious Diseases, volume 37, pages1081–1089(2018)
DOI: https://doi.org/10.1007/s10096-018-3220-z

*: Stefan A. Boers and Saskia D. Hiltemann contributed equally to this work.

Abstract

Microbiota profiling has the potential to greatly impact on routine clinical diagnostics by detecting DNA derived from live, fastidious, and dead bacterial cells present within clinical samples. Such results could potentially be used to benefit patients by influencing antibiotic prescribing practices or to generate new classical-based diagnostic methods, e.g., culture or PCR. However, technical flaws in 16S rRNA gene next-generation sequencing (NGS) protocols, together with the requirement for access to bioinformatics, currently hinder the introduction of microbiota analysis into clinical diagnostics. Here, we report on the development and evaluation of an "end-to-end" microbiota profiling platform (MYcrobiota), which combines our previously validated micelle PCR / NGS (micPCR / NGS) methodology with an easy-to-use, dedicated bioinformatics

pipeline. The newly designed bioinformatics pipeline processes micPCR/NGS data automatically and summarizes the results in interactive, but simple web reports. In order to explore the utility of MYcrobiota in clinical diagnostics, 47 clinical samples (40 "damaged skin" samples and 7 synovial fluids) were investigated using routine bacterial culture as comparator. MYcrobiota confirmed the presence of bacterial DNA in 37/37 culture-positive samples and detected bacterial taxa in 2/10 culture-negative samples. Moreover, 36/38 potentially relevant aerobic bacterial taxa and 3/3 mixtures of anaerobic bacteria were identified using culture and MYcrobiota, with the sensitivity and specificity being 95%. Interestingly, the majority of the 448 bacterial taxa identified using MYcrobiota were not identified using culture, which could potentially have an impact on clinical decision-making. Taken together, the development of MYcrobiota is a promising step towards the introduction of microbiota analysis into clinical diagnostic laboratories.

Introduction

The detection, identification, and further characterization of pathogenic microorganisms are the major step in establishing appropriate (antibiotic) treatment for infectious diseases. However, the causative microorganism of an infection may not always be detected using current "gold standard" culturing techniques. Further, most molecular-based detection methods, e.g., PCR, require a priori knowledge of the potential pathogen before a test is performed. To overcome these limitations, the bacterial composition can be defined and genera identified using a culture-free, broad-range PCR strategy that targets the prokaryotic 16S rRNA gene followed by next-generation sequencing (NGS) [0]. However, to date, 16S rRNA gene NGS methods to profile microbial compositions have been focused on research questions mostly, with only a few studies having evaluated the utility of 16S rRNA gene NGS methods for clinical microbiology [1, 2]. Currently, the utilization of 16S rRNA gene NGS methods within routine clinical diagnostics has been hindered by issues relating to the generation of PCR artifacts (e.g., chimera formation and PCR competition) and the susceptibility of 16S rRNA gene NGS methods to DNA contamination that is derived from the laboratory environment and/or the reagents/consumables used. These limitations hinder the standardization of current 16S rRNA gene NGS methods to such an extent that non-identical microbiota results may be obtained when repeatedly analyzing the same sample [3].

Recently, the authors published a micelle PCR/NGS (micPCR/NGS) methodology that limits the formation of chimeric sequences and prevents PCR competition via the clonal amplification of targeted 16S rRNA gene molecules [4]. In addition, the micPCR/NGS methodology allows for the utilization of an internal calibrator (IC) to calculate the number of 16S rRNA gene copies for each individual operational taxonomic unit (OTU) present within a (clinical) sample, which

conveniently enables the subtraction of contaminating bacterial DNA via the quantification of 16S rRNA gene copies within negative extraction control (NEC) samples. The authors showed that the microbiota results obtained using micPCR/NGS possess a much higher accuracy (precision and trueness) compared to those obtained using traditional 16S rRNA gene NGS protocols and that the ability to determine and subtract contaminating 16S rRNA gene copies, results in contamination-free quantitative microbiota profiles—with a limit of detection (LOD) of only 25 16S rRNA gene copies per OTU [5]. This low LOD allows for the detection of bacterial OTUs at very low abundances or can confirm the absence of 16S rRNA gene copies in culture-negative results. Based on these findings, the authors suggested that the micPCR/NGS protocol could possess distinct advantages when processing clinical samples for microbiota profiling compared to traditional (semi-quantitative) 16S rRNA gene NGS methods that remain vulnerable to false-positive results (e.g., chimeric sequences or contaminant DNA) and inaccurate measurements of the OTU relative abundances in polymicrobial clinical samples due to template-specific variations in PCR efficiencies (i.e., PCR competition). However, the analysis of 16S rRNA gene NGS data depends on the use of bioinformatics tools that are complex for non-bioinformatics educated technicians/clinicians to utilize, and the required bioinformatics skills are nowadays mostly absent in clinical diagnostic laboratories.

In this publication, we designed an easy-to-use bioinformatics pipeline to determine bacterial taxa from 16S rRNA gene sequences that together with the micPCR/NGS strategy are part of an "end-to-end" microbiota profiling platform (MYcrobiota). The bioinformatics pipeline enables the full analyses of the NGS data obtained, from raw sequence files to final web reports that summarize the quantitative microbiota results, without the knowledge of command line scripts that would normally be required by 16S rRNA gene NGS users. As a proof of principle, we explored the utility of MYcrobiota for use in the clinical diagnostic laboratory by processing a total of 47 clinical samples and then comparing the results to conventional "gold standard" culture results. The samples tested included 40 specimens that were obtained from a variety of damaged skin conditions for which a polymicrobial biomass was expected, and an additional 7 specimens, obtained from patients who were suspected of having (prosthetic) joint infections, for which a low bacterial biomass was expected.

Materials and Methods

Ethics statement

An acknowledged national ethics committee from the Netherlands (Medisch Ethische Toetsingscommissie Noord-Holland, `http://www.metc.nl`) approved the study protocol

(M015–021), and all experiments were performed on leftover material of the included clinical samples in accordance with the relevant guidelines and regulations. The national ethics committee waived the need for participant consent as all data were anonymized and analyzed retrospectively under code.

Sample collection and study design

This study was performed retrospectively using 47 clinical samples obtained from 47 subjects. The results obtained by routine bacterial culturing methods had been used to guide patient treatment and care. In this study, we re-analyzed these samples using MYcrobiota and compared the results to the initial outcome of the culture results. The 47 samples included in this study were derived from wounds (22), ulcers (10), abscesses (5), puss (1), erysipelas (1), erythema (1), and 7 synovial fluids obtained from patients with suspected (prosthetic) joint infections.

Routine bacterial culture

All samples were cultured according to standard laboratory protocols performed in our laboratory and stored at −80 °C for subsequent MYcrobiota analysis. The routine bacterial culture methods included a 48-h incubation at 35 °C on tryptic soy agar plates with 5% sheep blood (TSASB, Oxoid), colistin aztreonam blood agar plates (CAP, Oxoid), and cystine lactose electrolyte deficient agar plates (CLED, Oxoid) under aerobic conditions; a 48-h incubation at 35 °C on chocolate agar with Vitox supplement (CHOCV, Oxoid) under 5% CO_2 conditions; and a 48-h incubation at 35 °C on TSASB under anaerobic conditions. All Gram-negative rods, beta-hemolytic streptococci, *Staphylococcus aureus*, *Staphylococcus lugdunensis*, and anaerobic bacteria cultured were reported as potentially relevant bacteria, of which the identification of aerobic bacteria was obtained using MALDI-TOF mass spectrometry (Bruker). Note that in this study, we did not focus on optimizing culturing methods to increase the sensitivity of the culture results and the routine bacterial culture methods used may not be 100% efficient for culturing the bacteria that were detected with MYcrobiota.

Micelle PCR and NGS

DNA was extracted from all 47 samples using the High Pure PCR Template Preparation Kit (Roche) according to the manufacturer's instructions. In addition, DNA from the accompanying elution buffer was extracted as a NEC at the same time in order to allow the subtraction of contaminating bacterial DNA after NGS processing. The total number of 16S rRNA gene copies within each DNA extract was measured using a 16S rRNA gene quantitative PCR (qPCR)

according to Yang et al. [6], after which each DNA extract was normalized to contain either 10,000, or <1000 16S rRNA gene copies per microliter. A synthetic microbial community (SMC) sample, containing 10,000 16S rRNA gene copies of **Moraxella catarrhalis** (ATCC 25240), *Staphylococcus aureus* (ATCC 43300), *Haemophilus influenzae* (ATCC 10211), and *Clostridium perfringens* (ATCC 12915), was processed with each batch of clinical samples as a positive control (PC) sample. Prior to amplification by micPCR, 1000 or 100 16S rRNA gene copies of *Synechococcus* DNA were added respectively as IC to the normalized DNA extracts containing 10,000 or <1000 16S rRNA gene copies per microliter. One hundred 16S rRNA gene copies of Synechococcus DNA were also added to the NEC DNA extract. The IC was used to express the resulting OTUs as a measure of 16S rRNA gene copies by the use of a correction factor (sample OTU copies = sample OTU reads × (initial IC copies/IC OTU reads)) as previously validated elsewhere [5].

16S rRNA gene amplicon library preparation using micPCR was performed as previously published [5], but we utilized a different micPCR primer set that made it possible to replace the former Roche 454 NGS platform with the Illumina MiniSeq platform. In this study, micPCR amplification was performed using modified 515F (5′-TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG TGY CAG CMG CCG CGG TAA-3′) and 806R (5′-GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GGA CTA CNV GGG TWT CTA AT-3′) primers that amplified the V4 regions of 16S rRNA genes as recommended for Illumina NGS and which incorporated universal sequence tails at their 5′ ends to allow for a two-step amplification strategy. During the second round of amplification, dual indices and Illumina sequencing adapters were attached using the Nextera XT Index kit (Illumina). Paired-end sequencing of the 16S rRNA gene amplicon library was performed using the MiniSeq system in combination with the 2 × 150 bp MiniSeq System High-Output Kit (Illumina), after which FASTQ-formatted sequences were extracted from the MiniSeq machine for downstream analysis. We utilized the micPCR/NGS approach to process all samples, including the NEC and the PC, in triplicate in order to increase accuracy and to correct for contaminating bacteria DNA derived from the laboratory environment as previously described [6].

Bioinformatics pipeline

The bioinformatics pipeline designed during this study consists of 23 well-established mothur tools (v.1.36) [7] and an additional 9 custom-made tools developed by the authors that have been integrated and combined in Galaxy as a full analysis service to deliver 16S rRNA gene analysis for micPCR/NGS experiments. Essentially, we have incorporated the functionality of mothur in Galaxy, which is a project dedicated to simplify the use of complex command line bioinformatics

tools (such as mothur) using a user-friendly web interface [8, 9, 10], and added new calculator tools to allow for a completely automatic processing of quantitative micPCR/NGS data. Importantly, the bioinformatics pipeline presents the microbiota results together with an extensive overview of the quality control measurements performed during the micPCR/NGS data analysis, to the user in an organized fashion via an interactive web report. The complete workflow of the bioinformatics pipeline is visualized in Fig. 5.2. All the tools required for the bioinformatics pipeline can be found in Galaxy's Tool Shed (`https://toolshed.g2.bx.psu.edu/`). A workflow definition file can be downloaded from GitHub (`https://github.com/ErasmusMC-Bioinformatics/MYcrobiota`) and may be imported to any Galaxy platform, thereby offering the required set of bioinformatics tools. For more information on how to install and use this pipeline, please refer to the documentation in GitHub (`https://github.com/ErasmusMC-Bioinformatics/MYcrobiota`).

Quantitative PCR methods

The total bacterial biomass within each DNA extract was measured using a 16S rRNA gene quantitative PCR (qPCR) that targets the 16S rRNA gene V5-V7 region, which is a different region of the 16S rRNA gene compared to MYcrobiota [6]. Therefore, the 16S rRNA gene qPCR is a complementary technique that enables the validation of the MYcrobiota process when determining the total number of 16S rRNA gene copies. For this, CT values were related to a serial dilution of the previous calibrated and normalized SMC sample that contained mixed and equimolar concentrations of four bacterial species and ranged from a total of 100 to 10,000 16S rRNA gene copies per PCRi. In addition, the *S. aureus*-specific biomass was assessed within each DNA extract using a *S. aureus* qPCR that employs a *S. aureus*-specific marker as described by Martineau et al. [11]. Here, CT values were related to a serial dilution of only the calibrated *S. aureus* (ATCC 43300) DNA stock that ranged from a total of 10 to 10,000 copy numbers of the Martineau fragment. The PCRs were performed in 10-µL reaction volumes using the LightCycler 480 Probes Master (Roche) with the addition of 0.5 and 1.0 µM of each PCR primer for the 16S rRNA gene and *S. aureus* qPCRs respectively. Also, 0.25 µM of a Fam-labeled probe was added for the real-time detection of the 16S rRNA gene amplification, and 1× Resolight Dye (Roche) was added to the *S. aureus* qPCR in order to measure the *S. aureus* DNA amplification. All PCRs were performed using the following conditions: initial denaturation at 95 °C for 5 min followed by 45 cycles of PCR, with cycling conditions of 5 s at 95 °C, 10 s at 55 °C, and 30 s at 72 °C. Availability of data and materials

The datasets generated and analyzed during the current study are available in the Sequence Read Archive repository with accession number SRP109023, `https://www.ncbi.nlm.nih.gov/sra/?t`

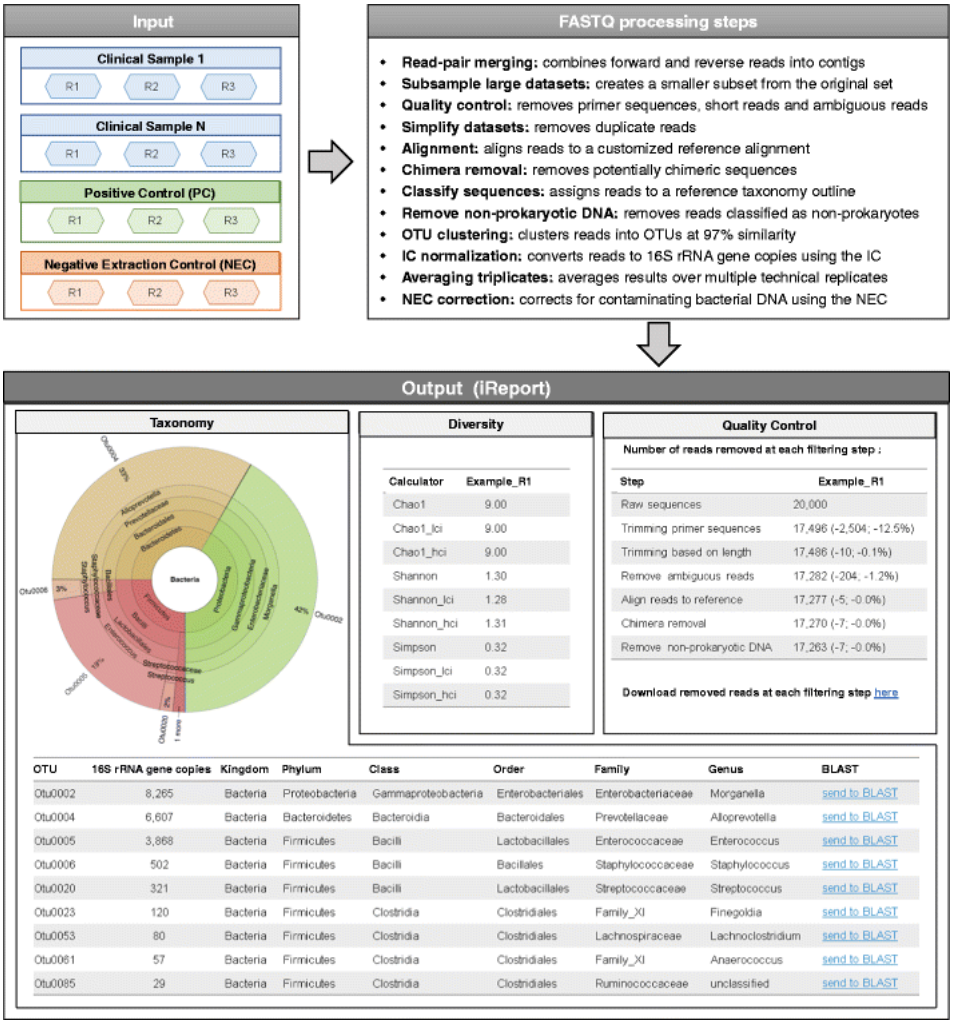


Figure 5.2: Schematical overview of the bioinformatics pipeline. FASTQ-formatted sequences obtained from triplicate experiments using micPCR/NGS (R1, R2, and R3) are automatically processed via the use of 32 (mothur) tools that have been integrated and combined in Galaxy as an "end-to-end" analysis service. The results obtained per sample (average of triplicate results) are presented to the user in a single, interactive iReport that consist of three tabs. The taxonomy tab visualizes and lists the resultant microbiota profiles. The diversity tab summarizes the results of three diversity calculators (Chao1, Shannon, and Simpson). The quality control tab provides an extensive overview of the quality control measurements during the analysis

erm=SRP109023.

Results

Development of an easy-to-use bioinformatics pipeline

In order to analyze 16S rRNA gene NGS data obtained using micPCR/NGS, we designed a Galaxy-based bioinformatics pipeline for use in clinical diagnostics. This workflow is largely based on the well-established standard operating procedure (SOP) defined by the creators of mothur [12]. We have adapted the SOP to our specific use-case by integrating several custom-made tools that allow for the subsampling of large datasets, the averaging over multiple technical replicates, converting the number of obtained sequence reads per OTU to 16S rRNA gene copies per OTU via the use of an IC, and correction for contaminating bacterial DNA via the use of NECs. All results are presented to the user as a single, interactive web report in Galaxy using the iReport tool [13]. The iReport was designed to visualize the resultant microbiota profiles using KRONA [14], list quantitative microbiota profiles in OTU tables (with the microbial load per OTU reported as 16S rRNA gene copies), summarize results of diversity calculators, and provide an extensive overview of the quality control measurements during the analysis. Importantly, the iReport is relatively small in size (6 MB per sample for our datasets) that enables easy sharing and storage of 16S rRNA gene NGS results (Fig. 5.2).

Validation of the MYcrobiota process

As shown in Fig. 5.3, MYcrobiota results obtained from the PC that was profiled in three independent experiments showed a median value of only a 1.3-fold (±0.2) difference between the measured 16S rRNA gene copies per bacterial species and the expected 10,000 16S rRNA gene copies per bacterial species present in the PC. In addition, comparisons between the measured 16S rRNA gene copies determined in actual clinical samples using MYcrobiota compared to qPCR results revealed an average of only a 1.5-fold (±0.5) and a 1.3-fold (±0.4) difference for the total bacterial biomass and the Staphylococcus OTU-specific biomass respectively. Of note, 10 of the 47 clinical samples included in this study resulted in culture-negative results, and the absence of bacterial DNA in these samples was confirmed with both qPCR and MYcrobiota methods. Also, one discrepant sample was detected that showed a 20-fold higher abundance of staphylococci detected by MYcrobiota compared to that detected by qPCR. This result can be explained by the presence of *S. aureus* and *S. non-aureus* within this sample. In fact, the *S. aureus* qPCR showed a 100% specificity compared to *S. aureus* culture-positive results and indicates the presence of *S. non-aureus* bacteria within 7 additional samples in which the *Staphylococcus* OTU was detected using MYcrobiota but no *S. aureus* could be cultured. Taken together, these data demonstrate the accuracy of the MYcrobiota process and the ability to incorporate quantitative results obtained from additional (species-specific) qPCRs.

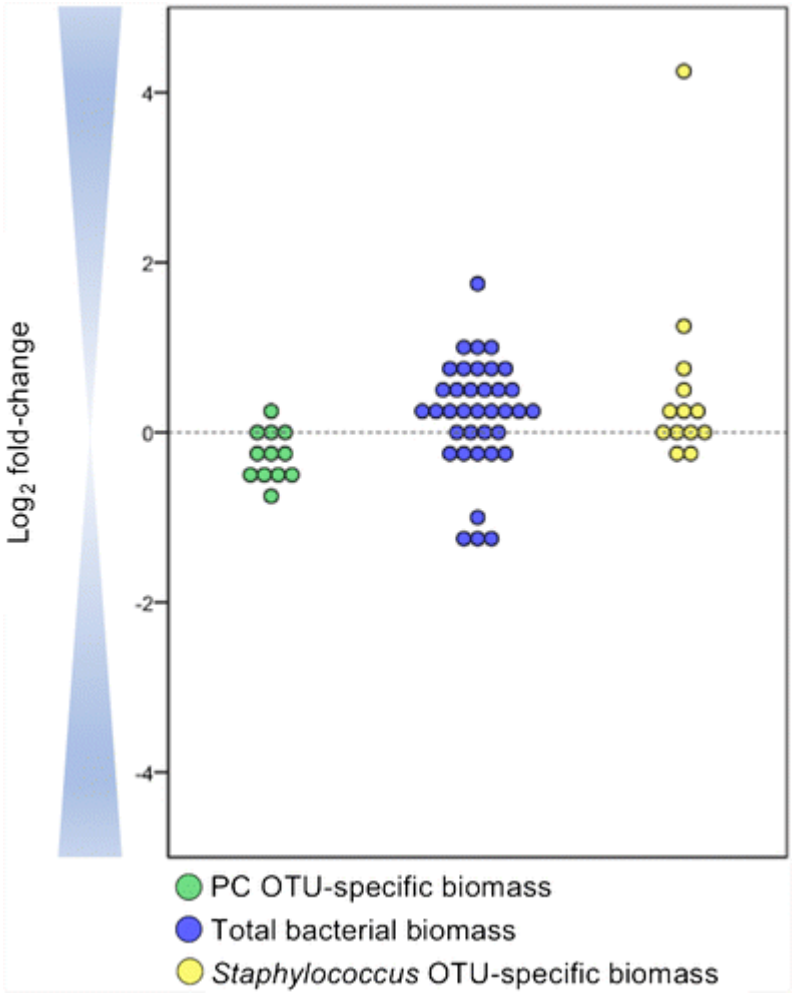


Figure 5.3: Accuracy of 16S rRNA gene copy determination using MYcrobiota. The expected number of 16S rRNA gene copies within the positive control (PC) was compared to the measured number of 16S rRNA gene copies using MYcrobiota (green dots). The PC contained 10,000 16S rRNA gene copies of four different bacterial species and was processed in three independent MYcrobiota experiments. The indirect estimation of the total bacterial biomass within 37 clinical samples using MYcrobiota was compared to the total 16S rRNA gene copies measured directly using a 16S rRNA gene qPCR (blue dots). The *Staphylococcus* OTU-specific biomass from 13 *S. aureus* culture-positive samples was compared to the *S. aureus* biomass detected directly using a *S. aureus*-specific qPCR (yellow dots). In order to compare the number of *S. aureus* genome copies estimated using qPCR to the number of 16S rRNA gene copies detected using MYcrobiota, the estimated *S. aureus* genome copies were first multiplied by a factor of 6 to correct for differences in copy numbers of the Martineau fragment and the 16S rRNA gene present on the *S. aureus* genome. The calculated differences between methods were plotted using a binary logarithmic scale

Comparing MYcrobiota results to routine bacterial culture

In order to explore the utility of MYcrobiota in the field of clinical diagnostics, we processed a total of 47 clinical samples and compared the results to routine bacterial culture. All bacterial genera

detected using culture and MYcrobiota are reported per sample in Fig. 5.4. Using standard bacterial culture techniques, our laboratory detected a total of 38 potentially relevant aerobic bacterial genera within 25 clinical samples and obtained a positive culture of a mixture of anaerobic bacteria in 3 samples. No bacteria were cultured from 10 samples, and an additional 10 samples resulted in the growth of bacteria that were all presumed to be commensal flora. In contrast, using MYcrobiota, we detected a total of 448 bacterial operational taxonomic units (OTUs) in 39 samples of which 337 OTUs (75%) could be identified as anaerobic bacterial genera that were detected in 21 samples. No bacterial DNA was measured in 8 out of 10 culture-negative samples. The sensitivity for bacterial culture detection by MYcrobiota was determined at 100% and the specificity at 83% using culture as "gold standard."

The majority of bacterial genera identified with culture were also identified using MYcrobiota. As shown in Fig. 5.0, MYcrobiota detected 36 of all 38 aerobic bacteria cultured on a genus-level taxonomy and confirmed the growth for anaerobic bacteria in 3 samples (sensitivity 95%; specificity 95%). Important to note, the two discrepant bacterial genera were measured using the micPCR/NGS strategy, but below the technique's LOD of 25 16S rRNA gene copies per OTU. In contrast, the vast majority of bacterial genera identified with MYcrobiota were presumed to belong to the commensal flora using culture or were not cultured at all (Table 5.4). These additional taxa include potential pathogens such as the *Kingella* OTU that was detected from a synovial fluid sample obtained from a juvenile patient that was not detected using culture and was confirmed using a *Kingella kingae*-specific PCR.

Discussion

In this study, we developed and explored the utility of an "end-to-end" microbiota profiling platform (MYcrobiota)—consisting of our previously published 16S rRNA gene sequencing methodology (micPCR/NGS) in combination with an easy-to-use bioinformatics pipeline—to investigate human samples for the clinical diagnostic laboratories. The bioinformatics pipeline designed during this study allows for a fully automated sequence interpretation of 16S rRNA gene NGS data that is obtained using the validated micPCR/NGS protocol without the need for advanced bioinformatics skills that are often unavailable in the clinical diagnostic laboratories. The MYcrobiota results are presented using (interactive) visualizations and tables, including an overview of all removed sequences during the analysis that allows for a manual evaluation of the quality measurements pre-installed within the bioinformatics pipeline. Moreover, connections of OTU representative sequences to the external NCBI database are available and can be used to ensure that the taxonomic identification of bacterial genera is correct [15]. Importantly, the

Sample	Routine bacterial culture	MYcrobiota
01_U	Commensal flora (1+)	Anaerobic bacteria (346,300), *Corynebacterium* (10,725)
02_U	Commensal flora (2+)	*Staphylococcus* (941)
03_W	Commensal flora (1+)	Anaerobic bacteria (263), *Streptococcus* (33), *Staphylococcus* (25)
04_U	*Pseudomonas* (3+), *Staphylococcus* (2+)	*Pseudomonas* (4,706), *Staphylococcus* (848), *Enterococcus* (135), Anaerobic bacteria (102)
05_U	*Proteus* (2+), *Enterobacteriaceae** (2+), *Streptococcus* (1+), Commensal flora (1+)	Anaerobic bacteria (8,271), *Proteus* (3,510), *Streptococcus* (632), *Enterobacteriaceae** (333)
06_U	Commensal flora (1+)	*Moraxella* (8,947), *Corynebacterium* (734)
07_W	*Enterobacteriaceae* (1+)	*Enterobacteriaceae** (5,386), *Bacillus* (44)
08_W	Negative	Negative
09_W	Commensal flora (1+)	Anaerobic bacteria (523), *Staphylococcus* (31)
10_A	Anaerobic bacteria (2+), *Pasteurella* (2+), *Streptococcus* (2+)	Anaerobic bacteria (3,704,750), *Pasteurella* (242,250), *Streptococcus* (28,625)
11_W	*Enterobacteriaceae** (3+), *Staphylococcus* (3+)	*Enterobacteriaceae** (3,420,786), *Acinetobacter* (1,126,632), *Staphylococcus* (32,760)
12_A	*Enterobacteriaceae** (2+), *Streptococcus* (2+)	*Enterobacteriaceae** (18,046), *Streptococcus* (6,409), *Enterococcus* (67)
13_Es	Commensal flora (1+)	*Staphylococcus* (344), Anaerobic bacteria (150), *Dermabacteraceae** (93), *Haemophilus* (64), *Corynebacterium* (53)
14_W	Commensal flora (1+)	*Staphylococcus* (31)
15_U	*Staphylococcus* (4+)	*Staphylococcus* (17,035)
16_U	*Enterobacteriaceae** (3+), *Stenotrophomonas* (2+), Commensal flora (2+), *Proteus* (1+), ***Pseudomonas* (1+)**	*Enterobacteriaceae** (828,310), *Proteus* (250,670), *Stenotrophomonas* (11,760)
17_W	*Staphylococcus* (1+), Commensal flora (1+)	*Staphylococcus* (4,886)
18_W	*Staphylococcus* (3+), Commensal flora (2+)	*Staphylococcus* (141,120), *Corynebacterium* (4,959)
19_W	*Streptococcus* (2+), *Staphylococcus* (1+)	*Streptococcus* (114,257), *Staphylococcus* (44,772), *Corynebacterium* (8,749), Anaerobic bacteria (897)
20_W	*Enterobacteriaceae** (3+), ***Staphylococcus* (2+)**	*Enterobacteriaceae** (4,574,310)
21_U	Commensal flora (2+)	*Moraxella* (1,066,608), *Acinetobacter* (142,155), *Pseudomonas* (30,051), Anaerobic bacteria (30,051), *Corynebacterium* (23,976), *Alkanindiges* (2,187)
22_W	*Staphylococcus* (2+), Commensal flora (1+)	*Staphylococcus* (105,648)
23_Et	*Staphylococcus* (2+), Commensal flora (2+)	*Staphylococcus* (14,803), *Corynebacterium* (66)
24_U	*Staphylococcus* (3+), *Streptococcus* (3+), Commensal flora (2+)	*Staphylococcus* (231,756), Anaerobic bacteria (96,740), *Streptococcus* (15,904), *Enterococcus* (1,680)
25_W	*Staphylococcus* (3+), Commensal flora (2+)	*Staphylococcus* (23,175), *Corynebacterium* (15,488), Anaerobic bacteria (1,271)
26_W	Commensal flora (1+)	*Staphylococcus* (4,142), Anaerobic bacteria (101), *Corynebacterium* (94), *Streptococcus* (47)
27_A	*Streptococcus* (1+)	Anaerobic bacteria (1,062,060), *Streptococcus* (5,490), *Treponema* (3,435), *Gemella* (1,425), *Mycoplasma* (870), *Tannerella* (720)

28_W	Commensal flora (2+), *Streptococcus* (2+)	**Anaerobic bacteria (114,004)**, *Streptococcus* (43,208)
29_A	*Streptococcus* (1+)	*Streptococcus* (12,225), **Anaerobic bacteria (3,384)**, ***Gemella* (299)**, ***Enterococcus* (295)**, ***Haemophilus* (221)**, ***Capnocytophaga* (156)**, ***Granulicatella* (122)**, ***Neisseria* (119)**, ***Rothia* (52)**, ***Lautropia* (35)**
30_W	Negative	Negative
31_U	*Acinetobacter* (2+), *Enterobacteriaceae** (2+), Commensal flora (2+)	*Acinetobacter* (518,396), *Stenotrophomonas* (423,320), *Enterobacteriaceae** (12,046), *Corynebacterium* (5,928), *Bordetella* (4,636), *Brevibacterium* (988)
32_W	*Staphylococcus* (2+), Commensal flora (1+)	**Anaerobic bacteria (251,692)**, *Streptococcus* (30,408), *Staphylococcus* (8,960)
33_W	*Staphylococcus* (1+), Commensal flora (1+)	*Staphylococcus* (466), **Anaerobic bacteria (171)**, *Streptococcus* (105), *Acinetobacter* (84), *Corynebacterium* (41)
34_W	*Staphylococcus* (3+)	*Staphylococcus* (218,141)
35_W	Commensal flora (2+)	*Staphylococcus* (5,121), **Anaerobic bacteria (769)**, *Roseomonas* (40)
36_W	Anaerobic bacteria (3+), Commensal flora (1+)	Anaerobic bacteria (493,183), *Streptococcus* (1,045)
37_W	*Streptococcus* (2+)	*Streptococcus* (11,457)
38_W	Anaerobic bacteria (3+)	Anaerobic bacteria (830,531)
39_P	*Streptococcus* (2+)	*Streptococcus* (10,277,376)
40_A	Negative	**Anaerobic bacteria (94,633)**, ***Enterobacteriaceae** (2,944)**, ***Streptococcus* (44)**, ***Thalassospira* (36)**
41_S	Negative	Negative
42_S	Negative	Negative
43_S	Negative	Negative
44_S	Negative	Negative
45_S	Negative	Negative
46_S	Negative	Negative
47_S	Negative	***Kingella* (25)**

Figure 5.4: Bacterial genera identified from 47 clinical samples using routine bacterial culture and MYcrobiota. Samples were derived from wounds (W), ulcers (U), abscesses (A), puss (P), erysipelas (Es), erythema (Et), and suspected joint infections (S). Cultured bacteria other than Gram-negative rods, beta-hemolytic *streptococci*, *S. aureus*, *S. lugdunensis*, and anaerobic bacteria were reported as commensal flora. The semi-quantitative culture results are presented as 1+, 2+, 3+, or 4+, depending on which quadrants demonstrate bacterial growth. The presence of anaerobic bacteria was reported as either a positive or a negative result. Bacterial species and OTUs detected using culture and MYcrobiota respectively are grouped at the genus level to compare results. Red shades indicate bacterial genera that were only identified by culture and blue shades indicate bacterial genera that were only identified by MYcrobiota (with "commensal flora" culture results representing a positive detection signal for any kind of aerobic bacterial OTU identified by MYcrobiota). The number of 16S rRNA genes measured using MYcrobiota is indicated in parentheses *Several bacterial genera that belong to the *Enterobacteriaceae* and *Dermabacteraceae* families could not be differentiated at a 97% similarity level using MYcrobiota

Bacterial taxa	Number of positive samples		Sensitivity (%)	Specificity (%)
	Routine bacterial culture	MYcrobiota		
Acinetobacter	1	4	100	98
Enterobacteriaceae*	7	8	100	98
Pasteurella	1	1	100	100
Proteus	2	2	100	100
Pseudomonas	2	2	67	100
Staphylococcus	14	20	93	100
Stenotrophomonas	1	2	100	100
Streptococcus	10	16	100	97
Anaerobic bacteria	3	21	100	71
Total	41	76	95	95

Table 5.0: Comparison of the cultured bacterial taxa to MYcrobiota results. The culture results are restricted to genus-level classifications in order to compare the OTUs detected using MYcrobiota to the culture-based results. The presence of anaerobic bacteria was reported as either a positive or a negative result. "Commensal flora" culture results were interpreted as a positive detection signal for any kind of aerobic bacterial OTU identified by MYcrobiota to perform specificity calculations
* Several bacterial genera that belong to the *Enterobacteriaceae* family could not be differentiated at a 97% similarity level using MYcrobiota

summarizing reports are relatively small in size, and storage of these files enables the traceability of patient test results that are required for clinical diagnostic laboratories according to quality requirements.

Using MYcrobiota, we processed a total of 47 clinical samples and compared the results to routine bacterial culture. Our results showed that the majority of bacteria identified with culture were also identified with MYcrobiota, but the majority of bacterial taxa identified with MYcrobiota were not identified using culture. Many of the additional bacterial taxa identified using MYcrobiota are obligate anaerobes that were commonly detected as a large component of the microbial population in samples obtained from damaged skin sites, which is consistent with previous studies [16, 17]. Indeed, it is well known that anaerobic bacteria are able to cause serious and life-threatening infections but are often overlooked due to their requirement for appropriate methods of collection, transportation, and cultivation [18]. Therefore, the culture-free MYcrobiota detection platform can play an important role in the identification of the bacteriological etiology of anaerobic infections or any other infections caused by fastidious microorganisms. Of note, it could be argued that the development of extensive culture techniques (so-called culturomics) may eventually facilitate the successful culture of supposedly "non-culturable" microbial isolates [19].

In addition to the accurate detection and identification of bacterial OTUs within clinical samples, MYcrobiota also provides the relative abundances in combination with the absolute abundances

for each detected bacterial OTU. This feature allows clinicians to obtain a comprehensive overview of the microbial composition of the clinical sample so that each quantified bacterial OTU, as well as the bacterial community as a whole, might be taken into account in clinical decision-making. Additionally, MYcrobiota allows for the removal of contaminating DNA from environmental sources in order to accurately and reliably investigate very low bacterial biomass, or no bacterial biomass, clinical samples [5]. For example, MYcrobiota confirmed the absence of 16S rRNA gene copies in eight of the ten samples that generated culture-negative results. The two discrepant samples contained either anaerobic bacteria or low amounts of the fastidious *Kingella* bacterium respectively. The ability to confirm culture-negative results improves the reliability of culture-negative diagnostic results. Additionally, the ability of MYcrobiota to detect bacterial OTUs at very low abundances makes MYcrobiota a suitable method to investigate normally sterile body sites, such as synovial fluids, cerebrospinal fluids, and blood samples. It should be noted however that the authors are aware of the fact that the construction of MYcrobiota is only a first step in the transition of microbiota research into actual clinical diagnostics. Extensive clinical and financial validation studies will be needed in order to validate and justify the routine introduction of molecular microbiota profiling methods into clinical diagnostic laboratories.

In conclusion, the stepwise development of MYcrobiota paves the way to introduce quantitative microbiota profiling into the clinical diagnostic laboratory. The method provides a highly accurate and comprehensive overview of the microbial composition of clinical samples or, alternatively, confirms the absence of 16S rRNA gene copies in culture-negative samples, using a standardized and validated 16S rRNA gene NGS workflow. Despite some shortcomings, e.g., lack of species identification and the inability to provide detailed information on antibiotic susceptibility, our data illustrates that MYcrobiota has promising applications in the field of clinical diagnostics and warrants investment in future studies to accurately evaluate the clinical relevance of 16S rRNA gene NGS results in clinical samples.

Notes

Compliance with ethical standards

An acknowledged national ethics committee from the Netherlands (Medisch Ethische Toetsingscommissie Noord-Holland, http://www.metc.nl) approved the study protocol (M015–021), and all experiments were performed on leftover material of the included clinical samples in accordance with the relevant guidelines and regulations. The national ethics committee waived the need for participant consent as all data were anonymized and analyzed retrospectively under code.

Conflict of interest

The authors declare that they have no conflict of interest.

Bibliography

[0] P.-E. Fournier and D. Raoult, "Prospects for the future using genomics and proteomics in clinical microbiology," *Annual review of microbiology*, vol. 65, pp. 169–188, 2011.

[1] D. D. Rhoads, R. D. Wolcott, Y. Sun, and S. E. Dowd, "Comparison of culture and molecular identification of bacteria in chronic wounds," *International journal of molecular sciences*, vol. 13, no. 3, pp. 2535–2550, 2012.

[2] S. J. Salipante, D. J. Sengupta, C. Rosenthal, G. Costa, J. Spangler, E. H. Sims, M. A. Jacobs, S. I. Miller, D. R. Hoogestraat, B. T. Cookson, *et al.*, "Rapid 16s rrna next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections," *PloS one*, vol. 8, no. 5, p. e65226, 2013.

[3] A. Hiergeist, U. Reischl, A. Gessner, *et al.*, "Multicenter quality assessment of 16s ribosomal dna-sequencing for microbiome analyses reveals high inter-center variability," *International Journal of Medical Microbiology*, vol. 306, no. 5, pp. 334–342, 2016.

[4] S. A. Boers, J. P. Hays, and R. Jansen, "Micelle pcr reduces chimera formation in 16s rrna profiling of complex microbial dna mixtures," *Scientific reports*, vol. 5, p. 14181, 2015.

[5] S. A. Boers, J. P. Hays, and R. Jansen, "Novel micelle pcr-based method for accurate, sensitive and quantitative microbiota profiling," *Scientific reports*, vol. 7, p. 45536, 2017.

[6] S. Yang, S. Lin, G. D. Kelen, T. C. Quinn, J. D. Dick, C. A. Gaydos, and R. E. Rothman, "Quantitative multiprobe pcr assay for simultaneous detection and identification to species level of bacterial pathogens," *Journal of clinical microbiology*, vol. 40, no. 9, pp. 3449–3454, 2002.

[7] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, *et al.*, "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and environmental microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[8] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko, "Galaxy: a platform for interactive large-scale genome analysis," *Genome Res*, vol. 15, no. 10, pp. 1451–1455, 2005.

[9] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor, "Galaxy: A web-based genome analysis tool for experimentalists," 2010.

[10] J. Goecks, A. Nekrutenko, and J. Taylor, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome biology*, vol. 11, no. 8, p. R86, 2010.

[11] F. Martineau, F. J. Picard, P. H. Roy, M. Ouellette, and M. G. Bergeron, "Species-specific and ubiquitous-dna-based assays for rapid identification of staphylococcus aureus," *Journal of clinical microbiology*, vol. 36, no. 3, pp. 618–623, 1998.

[12] J. J. Kozich, S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss, "Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform," *Applied and environmental microbiology*, vol. 79, no. 17, pp. 5112–5120, 2013.

[13] S. Hiltemann, Y. Hoogstrate, P. Van der Spek, G. Jenster, and A. Stubbs, "ireport: a generalised galaxy solution for integrated experimental reporting," *GigaScience*, vol. 3, no. 1, p. 19, 2014.

[14] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Interactive metagenomic visualization in a web browser," *BMC bioinformatics*, vol. 12, no. 1, p. 385, 2011.

[15] S. A. Boers, R. Jansen, and J. P. Hays, "Suddenly everyone is a microbiota specialist," *Clinical Microbiology and Infection*, vol. 22, no. 7, pp. 581–582, 2016.

[16] L. B. Price, C. M. Liu, J. H. Melendez, Y. M. Frankel, D. Engelthaler, M. Aziz, J. Bowers, R. Rattray, J. Ravel, C. Kingsley, *et al.*, "Community analysis of chronic wound bacteria using 16s rrna gene-based pyrosequencing: impact of diabetes and antibiotics on chronic wound microbiota," *PLoS One*, vol. 4, no. 7, p. e6462, 2009.

[17] K. Smith, A. Collier, E. M. Townsend, L. E. O'Donnell, A. M. Bal, J. Butcher, W. G. Mackay, G. Ramage, and C. Williams, "One step closer to understanding the role of bacteria in diabetic foot ulcers: characterising the microbiome of ulcers," *BMC microbiology*, vol. 16, no. 1, p. 54, 2016.

[18] I. Brook, "Clinical review: bacteremia caused by anaerobic bacteria in children," *Critical Care*, vol. 6, no. 3, p. 205, 2002.

[19] J.-C. Lagier, P. Hugon, S. Khelaifia, P.-E. Fournier, B. La Scola, and D. Raoult, "The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota," *Clinical microbiology reviews*, vol. 28, no. 1, pp. 237–264, 2015.

Scaling the Researcher

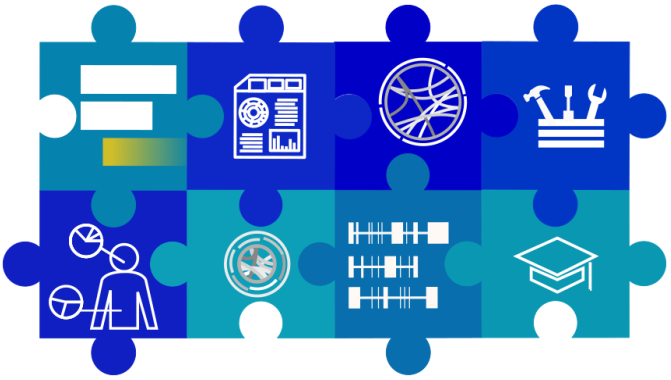
Meet Dana. Dana is a biologist who studies DNA. Dana sends her samples out to be sequenced, and gets back hard drives full of files. However, she doesn't really know what to do with these files. She tries opening a VCF file in Excel, but it complains that there are too many lines and immediately closes. She double clicks on a FASTQ file, and her computer freezes while trying to open the file in Notepad.

Luckily her group has a bioinformatician, Bindi, surely they can just do the analysis for her! Bindi tells her she is very busy right now doing analysis for the group, and can look at her files in a few months, hopefully. Dana just wants to analyse her data and publish her paper! She doesn't want to wait months, so she tries analyzing her data herself. She finds some tools in the literature but when she tries to install them, she sees they cannot be run on Windows, what now? She gets access to a Linux compute server, but the tool she wants does not install properly. There are no good instructions on how to fix the problem. She tries a different tool, which does install, but when she runs it on her own data, she gets some cryptic error messages. *"Only bioinformaticians could make sense of all this!"*, she sighs.

Then somebody tells her about Galaxy. It has the tools she wants, and she doesn't even have to install anything, all she has to do is click buttons in her browser! And it's free! Galaxy is not easy, but there are a lot of useful training materials available. She teaches herself how to use Galaxy, and how to analyze her data. She reads papers and reproduces their methods in Galaxy. She gets stuck, but asks for help from the community. She goes to workshops to learn more. It was a lot of work, but she gets some promising results! She has another batch of samples on the way. She creates a workflow to speed up the analysis next time.

Her coworkers also want to analyze their data. Bindi the bioinformatician is still very busy, and the queue is long, so they ask Dana for help. Dana shares her Galaxy workflow with them. She gets a lot of emails with questions from her coworkers, so she decides to create her own Galaxy tutorial about the workflow. Dana teaches a workhops, training her colleagues on how to use Galaxy and her workflow.

Dana publishes her paper. Dana is happy. With researchers in the group now running their own day-to-day analyses, Bindi the bioinformaticians has more time. She can use this time to add new tools to Galaxy, create workflows for the group, and develop training materials for her coworkers to use. Bindi is happy too.



"Humans are allergic to change. They love to say, 'We've always done it this way'. I try to fight that. That's why I have a clock on my wall that runs counter-clockwise."

Grace Hopper

6

Discussion

"Bioinformatics has become too central to biology to be left to specialist bioinformaticians" notes Lincoln Stein in 2008 [0]. And indeed, bioinformatics now plays a vital role in almost every

biological and biomedical research project. Data is being generated at an exponential rate, but bioinformaticians are not, and it is not scalable to leave the analysis of all this data to specialist bioinformaticians. Much of the discussion around the challenge of scaling to meet the needs of this data deluge tends to focus on the need for scaling up compute resources and data storage. And while these are both very important factors, the much greater challenge lies in scaling not just the analysis itself, but the interpretation of all this data. And nobody is better suited for this task than the research scientists themselves. So the question now becomes: *How can we scale the researcher?*

The answer to this question lies in empowering researchers to run their own data analyses again. The way we achieve this is by shining a light in the black box of bioinformatics. This entails making bioinformatics tools and pipelines accessible and easy to use for domain scientists. It requires training of researchers not just in how to use these tools, but also basic knowledge about the computational methodologies, and any biases these may introduce that can impact the interpretation of results.

This is not to say there is no role for bioinformaticians. It takes a lot of work to make analysis accessible, but by enabling research scientists to handle their own day-to-day analyses, we free up bioinformaticians to focus on the development of new and better tools, configuring new workflows, improving reproducibility, and developing and delivering training.

Therefore, in this thesis we set out to develop a user-friendly, open, and FAIR bioinformatics framework for NGS analysis, so that we can create a community of researchers who are computationally informed. We then applied this paradigm to a series of research projects to illustrate its utility.

6.0 Accessible Bioinformatics

The main challenge in our aim of scaling the researcher, is making bioinformatics accessible for non-bioinformaticians. This includes not only making analyses easy to use, but also easier to understand and interpret. In our approach we used the Galaxy platform to provide a user-friendly interface to bioinformatics analyses. Galaxy does all the heavy lifting of handling installation of the tools and all their dependencies, of optimally scheduling jobs across the compute resources, tracking provenance of analyses, enabling sharing, and much more. This leaves the researcher free to focus on the interpretation of their results. The second main component of our approach to making bioinformatics more accessible consists of providing extensive bioinformatics training. To this end, we founded the Galaxy training materials project, a collaborative framework for

development and maintenance of training materials using the Galaxy framework.

Galaxy as a user-friendly workflow platform

The Galaxy project [1, 2, 3] enables researchers to run complex data analysis without programming expertise, directly from their web browser. While there are several workflow management systems [4, 5, 6, 7, 8], some of these solutions are commercial and may not be affordable to many labs, and others require local installation, prohibiting scaling to accomodate large datasets. Furthermore, many solutions are not open source, meaning only the core development team can provide updates to the platform, limiting flexibility of the platform to its end users.

In contrast, Galaxy is completely free and open-source, and perhaps its most attractive asset is the very large and active user and developer community behind it. Galaxy encourages feedback and code contributions from its users to improve the platform, and evolve along with the ever-changing landscape that is bioinformatics. This paradigm means that anybody is able to add features to Galaxy or provide bug fixes, and can share those enhancements back to the main code base, thereby enabling a much faster and more flexible development cycle.

Galaxy continues to increase in popularity, illustrated by the exponential growth of both the number of available tools (over 8000) [9], and the number of publication citing Galaxy (almost 10,000) [10]. **Chapter 1** outlines the ongoing development in the Galaxy framework. One of the major improvements was in the handling of big data. Galaxy addresses this issue by introducing the concept of data collections. This allows users to easily scale their analyses up from single datasets to hundreds of thousands of samples at once, using the exact same procedures they have grown accustomed to. However, the greater challenge of big datasets is not just the processing of the files, but rather managing all this data in an effective way. To support the organisation of these large number of samples in Galaxy, the rule-based uploader was created, a novel feature which essentially allows sample sheets to be uploaded alongside the raw data files, and the metadata stored within them to be coupled to the datasets in Galaxy.

While the Galaxy project certainly greatly improves accessibility of bioinformatics analyses, there are still a number of areas of improvement remaining. For example, with sequencing now widespread and affordable, it has become feasible to use NGS directly in the clinic to inform patient care. However, until now Galaxy has focused mainly on researchers, and in order to make Galaxy more appealing for clinicians, the framework could benefit from a number of additional features. Firstly, a *locked-down*, workflow-centric user interface is required; in contrast with research applications where Galaxy's great flexibility is an asset, for clinical applications this

flexibility can be a hindrance, as it goes hand in hand with added complexity. Unlike researchers, clinicians do not typically need to play around with their datasets and try different tools; their analysis options should be limited to a small set of predefined and thoroughly validated analysis pipelines. While Galaxy does not (yet) offer such an alternate user view, it does offer API access to its framework, enabling custom front-ends to be developed on top of the Galaxy back-end. The need for such a simplified user interface to Galaxy is exemplified by the number of projects that have developed such a customized front-end to the Galaxy interface [11, 12, 13, 14], including several in-house projects at the ErasmusMC. For example, the IRIDA (Integrated Rapid Infectious Disease Analysis) project [12] is an open-source platform for public health genomics, which is currently in use by the Canadian public health agency to track infectious disease outbreaks. It uses Galaxy as its analysis platform, but provides a custom user interface tailored specifically to the needs of its users. This approach also enabled the project to integrate Galaxy directly with a third-party data management system where users can easily mange their projects and data. While Galaxy does not aim to be a data management system, in this big data era it has become increasingly imperative to offer good data management solutions to researchers. Therefore, integrations of Galaxy with third-party data management systems in the future could profoundly increase usability of the platform even further.

We used Galaxy as the main data analysis platform in each of the 3 research projects described in this thesis, including one clinical analysis platform (MYcrobiota). For several of these publications, we were able to submit our Galaxy history and workflows directly to the journal as a full end-to-end demonstration of our analysis pipeline, which could then be used to reproduce our results by reviewers and readers.

Visualisation and Reporting

Visualisation is an essential component to aid in interpretation of big and multidimensional datasets. Tools such as Circos [15] allow for the visualisation of multiple large and complex datasets in a single circular plot. Circos is highly customizable, but this high degree of flexibility comes paired with a high degree of complexity and a steep learning curve. To improve the user-friendliness of this tool and improve its interoperability with upstream tools, we developed Galactic Circos (Chapter 1), integrating Circos into Galaxy, and providing a set of preprocessing tools to support integration with a wide range of input file formats.

The Circos visualisation tool was especially useful in discovery of chromothripsis of the VCaP sample (Chapter 3). Here the highly rearranged nature of chromosome 5q was instantly obvious in one glance at the circular plot, in a way that simply inspecting the textual files listing the SVs could

not convey.

Effective visualisation of large data is crucial for the interpretation of analysis results, and in a broader sense, so is the reporting of results in a single, easily-digestible overview or report. Analysis pipelines often result in a large set of different output files, and displaying these results effectively to the end user tasked with interpretation of the results is a challenge in its own right. Galaxy in its current form lacks an appealing system of results summation and reporting. While Galaxy features a plugin system for visualisation of individual datasets, a generic reporting tool for displaying a set of workflow outputs together does not exist. To this end, we developed iReport (Chapter 1); a fully customizable Galaxy tool for the generation HTML reports capable of displaying any number of workflow outputs.

iReport is intended to be used as the final step of a workflow, and is generic enough that it makes no assumptions about the underlying analysis or scientific domain. iReport supports the ability to add links to datasets and external resources, create searchable and sortable tables, embed images and custom text, and to structure content into different pages. By adding such a summarizing report at the end of an analysis pipeline, clinicians and other end users are able to run workflows and view results with minimal instruction or knowledge of the Galaxy interface. We used iReport as the reporting tool for the MYcrobiota clinical analysis platform (Chapter 5), based on the requirements of the clinicians at the Streeklab Haarlem.

Since the development of the iReport tool, the need for a reporting system more tightly integrated with the Galaxy framework itself was recognized by the Galaxy core team, and they have since started to integrate similar functionality into the Galaxy code base. While this reporting system is relatively new and does not yet support the full functionality of iReport, we are hopeful that continued development here will improve Galaxy's reporting mechanisms.

Training

Training is an essential component in the dissemination of accessible bioinformatics tools and workflows. The Galaxy platform is especially well-suited for the delivery of bioinformatics training because it provides a layer of abstraction that allows trainees to focus on the bioinformatics *concepts* rather than the implementation details of the tools. Without this separation, trainees would have to simultaneously learn about the UNIX command line or programming environment, on top of the bioinformatics topics at hand. This would increase the cognitive load and hamper the learning process [16]. Given this observation of Galaxy's suitability for use in training, the Galaxy Training Network (GTN) [17] was formed; a loosely-defined open group of instructors around the world

who use Galaxy for training purposes. Initially there was little coordination between the different instructors in terms of materials used, and thus a lot of duplication of effort. There was a clear need for centralisation of training materials and knowledge sharing within the trainer community. Chapter 2 describes the community-driven web-based framework for the delivery of bioinformatics training using the Galaxy platform that we developed in response to this need. Our aim was to create a fully open and transparent framework that is accessible and easy to use for both trainees and trainers. The materials are centered around *research stories*; usually the recreation of results described in published papers. This gives trainees the confidence that the tools and pipelines are practically useful and of publication-level quality, as well as providing them with the opportunity to dive deeper into the science and informatics behind the training. Since the creation of this training platform, a number of scientific publications have included Galaxy training materials as a form of documentation and illustration of the presented analysis pipelines [18, 19, 20, 21].

One of the main challenges in designing this framework was to allow easy contributions from instructors, without the need for any web development knowledge. To this end, we used Jekyll templating [22], which allows tutorials to be written in the simple and accessible markup language called Markdown [23] which can be rendered as HTML. Analogous to how Galaxy allows scientists to run analyses while being abstracted away from the implementation layer of the tools, this approach allows instructors to create web pages for their tutorials without being concerned with the syntax and intricacies of the web application layer.

A further challenge was to enable the materials to be usable both by instructors during workshops, and by individuals learning on their own. This is accomplished by including all materials instructors might provide during a workshop in the GTN training materials framework. This includes introduction slides as well as hand-on materials, input datasets and workflows, further reading suggestions, and an automatically updated list of available Galaxy servers which meet the requirements to run a given tutorial. Furthermore, learning assessments are provided in the form of question boxes, answers to which are included within the materials (in an initially hidden state) for trainees to verify their understanding of the materials. If further assistance is required, links to support channels such as a help forum and chat rooms are provided.

The community-driven nature of the training framework is essential for the long-term survival of the project; it allows for the distribution of the maintenance burden and takes advantage of the combined expertise present in the community. All development happens on GitHub [24], where anybody may suggest additions or changes, and any such proposed changes are thoroughly tested using the Travis continuous integration system [25] to ensure functionality and adherence

to guidelines. The proposed changes are subsequently reviewed by one or more of the dedicated topic maintainers, or other volunteers from the community. Once approved, the code is merged into the main code base, and the new website is automatically built and deployed using Travis and GitHub. In order to assess the quality of the tutorials and identify areas of improvement, feedback from both trainees and instructors is indispensable. To this end, we integrated evaluation forms at the end of each tutorial, and hold regular community meetings with tutorial authors and trainers.

As Galaxy evolves, so will the associated tutorials; where Galaxy is expanding beyond bioinformatics and is now also being used in fields such as natural language processing and computational chemistry, so have we noticed a steady expansion of topics and tutorials contributed by the community. In the year following the publication of Chapter 2 , we saw 6 new topics added, 66 new tutorials, and the number of contributors grew from 64 to 137.

While the focus of development in this project initially lay with improving the experience for end-users of the tutorials, our focus is now shifting to increasing support for tutorial contributors and instructors intending to use our materials. The main challenge in the coming years will be the community management; creating and sustaining a close-knit community of Galaxy users and instructors so that the project can survive even when its original developers have moved on.

While this project focuses on the training of research scientists to effectively use bioinformatics analyses, the reverse is also important; much like bioinformatics may feel like a *black box* to researchers and clinicians, so can the wetlab seem like a black box to many bioinformaticians. Bioinformaticians typically work on multiple projects at the same time, often covering a variety of different scientific domains, and therefore can not acquire the same level of knowledge about the underlying biology as the domain specialists. However, similar to how researchers and clinicians will benefit from a basic knowledge of computational concepts to aid the interpretation of analysis results, so do bioinformaticians need a minimum knowledge of the scientific domains in order to optimally develop their tools and workflows. The tutorials in the GTN framework are generally aimed at novices and as a result can also be used by bioinformaticians to acquire knowledge of the basic concept involved in different scientific domains. However, I would love to see a project conceptually similar to the GTN training materials framework, but explicitly aimed at training bioinformaticians in the relevant biological concepts.

6.1 Use Cases

The concepts and tools described in the previous sections were applied to two separate use cases. Chapters 3 and 4 describe the creation of analysis tools and pipelines for variant analysis in prostate cancer research. In Chapter 5, Galaxy-based analysis pipelines were developed and tested for the application of NGS-based microbiota profiling for clinical diagnostics.

Cancer Analysis: Structural Variant Analysis

The Bio

Cancer is a disease of the genome, where DNA mutations accumulate over time and wreak havoc on the cell. Our best hope in defeating this disease lies in the full characterization of the cancer cell, not only in terms of their observed mutational signatures, but also the mechanisms behind these mutational landscapes.

While small-scale mutations that affect the protein-coding regions of the genome have been widely studied, the impact of non-coding mutations and large-scale structural variants (SVs) in cancer remains largely undefined [26, 27]. Now that whole-genome sequencing has become more available and affordable, a comprehensive characterization of SVs has become possible.

Large-scale genomic rearrangements have the potential to lead to the generation of hybrid genes known as fusion genes. Accurate detection of such fusion genes may aid in the diagnosis or treatment of cancer [28, 29, 30, 31]. Prostate cancer was among the first solid tumours demonstrated to harbour frequent large-scale genomic rearrangements, demonstrated by the discovery of the *TMPRSS2-ERG* fusion gene, which was found to be present in approximately 50% of all prostate cancers [32]. Subsequent studies have shown that point-mutations occur relatively infrequently in prostate cancer as compared to structural variations and extensive copy-number variation, suggesting that it is these large-scale genomic rearrangements that are the primary driver of prostate cancer progression [33, 34].

Understanding the mechanisms behind the acquisition of these structural variations will provide valuable insights into tumor progression. While historically the acquisition of mutations in tumour cells has been thought of as a gradual and incremental process, subsequent insights have uncovered several mutational processes capable of generating large numbers of mutations in a single event [35, 36]. Kataegis (from the greek word for thunderstorm) represents a *mutation storm* of localised hypermutations, and has been observed in multiple cancer types [37, 38]. Chromothripsis is a shattering of the genome –often an entire chromosome or chromosome arm– followed by a highly

erroneous repair step [39, 40]. Chromoplexy (from the Greek *pleko*, which means to weave) is a phenomenon where large chains of rearrangements affecting multiple chromosomes occur [41]. All these mechanisms challenge the classical view of cancer progression as a gradual and step-wise process.

In Chapter 3 we showed that chromothripsis also occurs in prostate cancer, by describing the identification of chromothripsis in the 5q arm of the VCaP cell line. A very large number of SVs were detected, and copy number varied primarily between two copy number states, with only sporadic occurrences of a third state. This alternation between a small number of copy number states suggest the rearrangements were precipitated by a single catastrophic event. Other studies have also observed chromothripsis in other prostate cancer samples [42, 43].

Such a high degree of genomic rearrangement has the potential of contributing to cancer progression e.g. through activation of oncogenes, or inactivation of tumor suppressor genes. To investigate this further, we evaluated the 573 rearrangements involving the impacted region of chromosome 5 for their potential to lead to the formation of fusion genes, using the iFUSE application presented in Chapter 3. Out of this large number of rearrangements, a relatively small number (18) were found to occur between two different genes at a consistent orientation, with only 2 predicted to be in-frame by the iFUSE application. Out of the 18 fusion candidates identified, 16 were confirmed on the DNA level using custom PCR. Only 5 of these were also measured on the mRNA level, suggesting instability of the fusion transcripts or down-regulation of the expression of these fusion genes. Therefore, our results suggests that chromothripsis does not preferentially impact coding regions and that there is no positive selection for in-frame fusion transcripts.

Overall, our research has shown that any studies involving this commonly used prostate cancer cell line should take the presence of chromothripsis on chromosome 5q into account. Furthermore, this study highlights the potential utility of this cell line as a model for research on chromothripsis.

Chromothripsis has been observed in many other cancer types as well, with the frequency of occurrence varying greatly across tumour types [35, 44, 45, 46]. A study examining 2,658 WGS patient samples obtained from the ICGC and TCGA projects found pervasive chromothripsis across all 38 cancer types examined. The prevalace varied significantly by tumour type, with liposarcomas (100%) and osteosarcomas (77%) on the high end, and thyroid adenocarcinomas (3.3%) and chronic lymphocytic leukemia (1.2%) on the lower end of the spectrum [35]. It must be noted that due to the variability in detection methods and in the precise definition and criteria of chromothripsis used in different studies, the exact numbers vary significantly in the literature, though the variability across different tumour types is consistently observed. Chromothripsis has

been shown to be associated with poorer outcome [47, 48, 49, 50] and accurate detection of the presence of chromothripsis could therefore provide clinically relevant prognostic value.

The Informatics

Whole-genome sequence experiments typically generate very large output files, and interpretation is greatly aided by visualisation and annotation with external data resources. Furthermore, in the case of structural variant analysis, the output file formats lack consistency across tools and sequencing platforms. To this end, the iFUSE application was developed; it supports multiple input formats, and creates a visual representation of fusion gene candidates and computes various metrics such as predictions of fusion protein product.

While iFUSE provides valuable aid in interpretation on a per-event basis through its visualisation and annotation, it is less suitable for obtaining a genome-wide overview of structural rearrangements. Large-scale rearrangements such as chromothripsis for instance are easy to miss when examining the raw textual output files or the iFUSE visualisations. In order to evaluate the presence of chromothripsis, a more high-level view of the genome is needed, and to that end, whole-genome visualisations were created using Circos [15] for rearrangements, and GNUplot [51] for the copy number and heterozygosity plots. This approach enabled the instant identification of chromothripsis present in the VCaP sample, which could then be followed up by closer examination of individual rearrangements involving the chromosome 5q using iFUSE.

Both the iFUSE application, the Circos Galaxy wrappers, and the other tools and visualisations created in this chapter are open-source and publicly available on an example server, and are accompanied by extensive documentation to enable re-use by institutes that may have legal restrictions about the use of clinical samples on public servers.

If we want to go beyond the mere detection of chromothripsis to the prediction of the effects of these catastrophic events on the cell, we will have to fully reconstruct the tumour genome. While algorithms have been developed for the reconstruction of structurally rearranged genomes [52, 43], none of these methods is currently capable of handling the high degree of rearrangement observed in some cancers, let alone those harbouring chromothripsis. This appears to be a limitation of the power of the data yielded by short-read techniques, rather than a shortcoming of the algorithmic methods. The recent advances in long-read sequencing techniques will provide valuable improvements for reconstructing cancer genomes impacted by chromothripsis by providing reads spanning multiple breakpoints. Indeed, this method was employed recently for the successful reconstruction of a genome impacted chromothripsis in a patient with the rare genetic Langer–Giedion syndrome [53]. Similar approaches could be used to reconstruct highly

rearranged tumour genomes and have the potential to significantly increase biological insight into the effects of chromothripsis events on the cell.

Cancer Analysis: Somatic Mutation Determination

The Bio

One of the main challenges in analysis of oncological samples is the identification of those mutations that potentially function as oncogenic drivers, and the large number of passenger mutations or polymorphisms present in the germline of the patient that are typically deemed to be functionally benign [54]. To this end, a sample of normal tissue from the same individual is often sequenced in conjunction with the tumour sample. This allows the subtraction of germline variants from the set of mutations detected in the cancer sample, in order to narrow down the set of potential driver mutations. However, in practice such an associated normal sample may not always be available, for a variety of reasons. In such cases, an alternative approach is required.

Given the observation that the majority of any individuals germline variants are polymorphic and observed frequently throughout the human population [55, 56], in combination with the exponential increase in publicly available genomic datasets, we explored the feasibility of constructing a so-called *virtual normal*, consisting of a reference set of variants found in the population. Chapter 4 describes this investigation.

Our approach used a set of over 900 publicly available whole genomes from healthy, ethnically diverse individuals. We combined this with the customary approach of annotating variants for presences in several online databases of polymorphism in the human population. We tested the performance of our method using 4 different tumour samples with associated normal samples, 2 of which had been sequenced on two different platforms (Complete Genomics and Illumina).

Our results show that while highly unique personal variants cannot be 100% corrected for, a significant number of variants observed in the tumour sample also occur in the virtual normal set of genomes, and thus can be corrected without the need of an associated normal sample from the same individual. As more and more WGS samples are made publicly available, increasingly rare variants may be corrected for in this manner. Furthermore, we demonstrated that the use of a virtual normal provided a significant improvement over relying solely on annotation with online variant databases. This observation seems counter-intuitive given that these databases often contain variants originating from tens of thousands of sequencing experiments; far more genomes than the virtual normal. However, this result can be explained by noting that the virtual normal approach preserves the genomic context of variants (e.g. adjacent variants in the

same sample), while this contextual information is lost when variants are submitted to variant databases. To understand why this contextual information is so important, one must realize that many mutations can be described in multiple different yet equivalent ways, and the only way to resolve this equivalency is to take the genomic neighbourhood of the variants into account. This enables more advanced comparison algorithms to resolve equivalency of variants where routine position-based exact comparison method fall short.

A combination of all 3 correction methods (matched normal, virtual normal, and variant databases) yields optimal results, but when faced with a choice between a virtual normal or a matched normal, both approaches performed roughly equally well. Without an associated normal, personal germline variants may be erroneously deemed somatic, but conversely, omission of the virtual normal led to a roughly equal number of false-positive somatic variants. This can be explained by a combination of sequencing errors or suboptimal variant calls in the associated normal sample, and the general difficulty present in variant comparison analyses.

Depending on the use case, the decrease in power to detect highly personal germline variants may be offset by the decrease in cost from the absence of the necessity of sequencing a matched normal sample with every tumour sample.

The Informatics

The samples in this study were sequenced by Complete Genomics [57], a sequencing service which delivers both raw sequencing data and post-processed results. Complete Genomics provide a suite of command line tools for handling and downstream analysis of their often custom file formats. As a first step towards building the virtual normal analysis pipelines, these existing tools were wrapped into Galaxy. On top of these third party tools, several custom analysis components had to be created, as well as several file format conversion steps to function as a *glue* between steps. The full suite of analysis tools has been made publicly available on GitHub, as well as detailed instructions on where to obtain the virtual normal genome set, and how this may be extended with additional normal samples in the future.

Microbiota Profiling for Clinical Diagnostics

The Bio

The MYcrobiota project described in Chapter 5 was aimed at developing a 16S microbiota profiling pipeline suitable for use in clinical diagnostics. The MYcrobiota platform is the result of a close collaboration with Streeklab Haarlem [58], a microbial diagnostics lab servicing a large number of GPs and hospitals in the region.

While 16S rRNA sequencing is a relatively well-established technique, there are several obstacles to overcome to enable its use in routine diagnostics. These obstacles include 1) the high prevalence of chimera formation during PCR amplification [59], 2) the inability to standardize the relative abundance results obtained from 16S profiling across different studies, and 3) the lack of a user-friendly bioinformatics pipeline that can be operated by clinicians without extensive bioinformatics knowledge. The MYcrobiota platform addresses obstacles; the first obstacle is overcome by using a novel PCR method, and the latter two obstacles are overcome through the creation of a Galaxy-based FAIR analysis platform adhering to the bioinformatics best practices outlined in this thesis.

In order to facilitate the use of 16S rRNA sequencing in a diagnostic setting, several enhancements to standard procedure were required, both in the wet lab and the bioinformatics pipelines to overcome the aforementioned obstacles. While chimera formation can be detected using in-silico methods, after which those sequences are typically removed from the dataset, it would of course be preferable to prevent their formation altogether. The MYcrobiota pipeline utilizes the micelle PCR (micPCR) method [60, 61]. In this approach, PCR amplifications of each 16S template sequence occurs in a physically distinct reaction environment called a micelle, thus greatly reducing the generation of hybrid sequences known as chimeras. The in-silico methods for chimera detection are still included in the workflow, and can now serve the additional purpose of providing quality control metrics for the effectiveness of the micelle method of a given run. Traditionally, 16S rRNA sequencing can only be used to provide relative abundance information, due to PCR competition induced bias. However, by utilizing the micelle PCR method, this bias is eliminated, allowing absolute quantification of abundance, enabling us to provide additional clinically relevant information in the MYcrobiota platform.

When dealing with clinical pipelines, validation is crucial. This validation needs to occur at multiple levels; each of the tools needs to be verified to work as expected, the interplay between tools when run together in a workflow needs to be verified, and of course validation on the experimental level. The Galaxy platform provides a framework for incorporating technical validation tests into the tool definitions themselves; by utilizing Galaxy in our application, we automatically gain the benefit of these technical validation tests. In a similar fashion, Galaxy allows specification of workflow-level validation tests, allowing us to validate the interplay of the different tools in our workflow. Aside from these technical validation steps, experimental validation must be part of the analysis protocol. Therefore, within the MYcrobiota standard, each sample is sequenced in triplicate and averaged in order to eliminate any remaining quantification bias in the micPCR protocol. Furthermore, a negative control sample is included in each batch, and finally, an internal

calibrator (IC) was used to enable quantification of each resulting OTU in terms of the number of gene copies rather than relative abundances. This IC consisted of a known quantity of a bacterium not present in the natural microbial flora under investigation, and was added to the samples before PCR amplification. By incorporation all these validation methods in the MYcrobiota application, we increase confidence in the results of the analyses, and allow us to more quickly identify potential problems when they arise.

The MYcrobiota approach was evaluated through the analysis of 47 clinical samples obtained from patients presenting with a variety of damaged skin conditions, and results were compared to the culture-based methods currently employed for routine clinical microbial diagnostics. The results showed that the vast majority (>95%) of genera detected by routine culturing were also detected by the MYcrobiota platform. Conversely, the majority of bacterial taxa detected by MYcrobiota were not identified by culture. Many of these additional genera detected were anaerobes consistent with previous studies [62], and included potential pathogens such as *Kingella* not detected in routine culture. The universality of the MYcrobiota pipeline has been subsequently demonstrated through its application to environmental studies in drinking water distribution systems [63] and in a clinical setting involving patients presenting with suspected septic arthritis [64].

It must be noted however, that certain limitations to the MYcrobiota remain, and further development and extensive clinical validation studies are required before introduction into routine diagnostics. For example, the current methodologies lack the discriminative power to differentiate to the species taxonomic level, which is often essential for clinical diagnostics. However, results could be supplemented with species-specific PCRs. Alternatively, relatively simple alterations to the current MYcrobiota platform could accommodate approaches such as the sequencing of multiple hypervariable regions, or even the full-gene 16S rRNA, as well as other potential genetic markers such as *rpoB* [65], *gyrB* [66], the ITS region [67], or one of many other potential markers capable of differentiation of prokaryotes at the species taxonomic level [68, 69]. As the potential for using 16S rRNA sequencing in the clinic becomes more apparent, an increasing number of studies have investigated how to further optimize clinical value of these methods [70, 71], further facilitating the incorporation of NGS analyses in clinical diagnostics.

In conclusion, the development of the MYcrobiota platform paves the way for the introduction of culture-free quantitative microbiota profiling methods into clinical diagnostic laboratories. It is capable of providing a highly accurate and comprehensive profile of the microbial composition of clinical samples, even at low biomass, and may provide clinicians with valuable information on potential pathogens not (easily) provided by the standard culture-based methods. Alternatively,

the culture-negative status of clinical samples may be confirmed by the absence of 16S rRNA gene copies in the MYcrobiota results. By providing an accessible platform such as MYcrobiota, which can be used with minimal bioinformatics expertise, we pave the way for standardisation between different labs, improving overall patient care.

The Informatics

Tools. As a first step to creating the MYcrobiota platform, we incorporated the full suite of 125+ mothur tools [72] into Galaxy, as well as the Krona [73] tool, and Phinch [74] display application for visualisation of results. To facilitate the interoperability of these tools and others already available in Galaxy, we also created some file format conversion tools to function as the *glue* between steps and facilitate the interoperability with existing downstream tools through the support of widely used file formats such as the BIOM format [75].

Workflows. In order to optimize utility of the tools, we provided several standard pipelines in Galaxy, based on available standard operating procedures (SOPs) defined in the research community. However, for use in the clinic, we made further customizations to tailor the workflows to support the specific experimental setup employed by Streeklab Haarlem. This included providing support for replicates, negative extraction controls, and the internal calibrator (IC) used for quantification. To accommodate these custom requirements, the standard pipelines had to be augmented with additional components and custom parameter settings, arrived at through a lengthy cycle of testing and adjustment, followed by clinical validation. In order to facilitate scaling of these analyses, the pipelines utilize Galaxy collections, enabling analysis sizes from single samples to tens of thousands of samples.

Because this workflow was intended for use directly in the clinic, extensive validation of the tools and workflows was performed, in collaboration with the Streeklab Haarlem. This included testing of the individual workflow components (tools), as well as the entire end-to-end analysis pipeline. Testing was performed using public datasets, artificially constructed samples, and previously analyzed patient samples. By analyzing mock communities, we were able to test the entire experimental pipeline, from sequencing to bioinformatics, and create and estimation of the error rates. In the final report at the end of the workflow, a set of QC metrics is presented to the user, to aid in the quick identification of potential problems with the sequencing or subsequent analysis, by flagging metrics outside the range of expected values based on the validation process.

Furthermore, privacy of datasets is always a primary concern in clinical bioinformatics. Depending on the type of data and experimental design, there may be legal and ethical restrictions on data transfer. Therefore, the datasets were decoupled from their clinical metadata, and anonymized

before entering the Galaxy pipeline. Furthermore, we also created a Docker [76] image of the entire MYcrobiota platform. This allows clinics to run MYcrobiota in-house with relatively little effort, offering additional flexibility and eliminating the need for data transfer.

Visualisation and Reporting. The MYcrobiota pipelines generate hundreds of files per run per sample, so a tailor-made web report was configured using the iReport tool described in Chapter 1 to aid clinicians in the interpretation of results. This report included several integrations with external resources such as BLAST and prokaryotic databases.

Open science and bioinformatics best practices. In order to optimize the utility of the tools and pipelines both now and in the future, all components of the MYcrobiota are open-source and publicly available on GitHub, and under testing using the Travis continuous integration platform. The advantages of this approach are illustrated by the fact that since its release, the tools have received numerous updates and bug fixes from members of the Galaxy community, relieving us, the original authors, of the long term maintenance burden and keeping the tools relevant as the underlying mothur components evolve.

Training. Galaxy training materials were developed to facilitate the dissemination of the MYcrobiota tools and pipelines. These were integrated into the training infrastructure developed in Chapter 2, and have since been used in numerous workshops by a variety of instructors in the Galaxy Training Network, as well as for self-study by individual learners online. Due to the feedback mechanisms built into the Galaxy training framework, learner feedback is collected, and we have received and implemented many suggestions for improvements, enabling incremental development and refinement of the materials.

6.2 Futuromics: Future Perspectives

The field of bioinformatics is in a constant state of flux, with data being generated at an exponential rate, and new sequencing technologies ever on the horizon promising greater biological insight. Many of the currently used tools, algorithms, and file formats will evolve or be replaced by new ones, and the challenge will be to create and maintain the IT infrastructure required to support these novel tools and techniques and to scale with the exponential rate of data analysis in this era of *big data*.

Sequencing

As sequencing technologies continue to evolve, so will the entire ecosystem of bioinformatics tools and algorithms surrounding them. The majority of microbiota profiling has been limited to a single hypervariable region of the 16S rRNA gene, but as long-read technologies mature, whole-gene sequencing of the 16S gene may increase resolution and allow for taxonomic determination down to the species level, which will greatly increase its utility in clinical applications [77, 78]. Long-read technologies are equally promising in cancer analyses, where they have the potential to span large-scale structural variants and greatly aid in the reconstruction of highly rearranged tumour genomes or even chromosomes affected by chromothripsis [79, 80]. Currently, such technologies are limited due to a high error rate as compared to traditional short-read sequencing technologies, but these error rates are quickly approaching competitive levels [81]. A further challenge is the limited availability of tools in the community aimed at the use of such relatively novel technologies, which will be remedied by the simple passage of time, and as the technologies improve, so does the incentive to develop tools capable of analysing the resulting data.

Similarly, the advent of single-cell sequencing technologies has the potential to greatly aid cancer genetics research by providing an accurate view of the state of a single cell within a tumour. This allows for the characterization intratumor heterogeneity by sampling physically distinct locations, as well as for the exploration of the role of rare cells in tumor progression [82].

Reference genome

The original human reference genome was represented as a single, linear, haploid sequence and was based on the DNA from a small number of individuals [83]. While unequivocally a transformative and landmark achievement, the structure of the reference genome comes with a set of limitations. Each individual, on average, has ~4 million small variants and around 2500 structural variations compared to the reference genome [84, 85]. This high degree of variability may result in difficulties in mapping and variant calling. Furthermore, because it is based on a relatively small number of subjects, it does not represent the most common allele in all locations, and even harbors some potentially pathogenic alleles [86, 87, 88]. Indeed, large-scale projects like the 1000 Genomes project [55] have analyzed large and ethnically diverse cohorts, and have revealed that the hg19 reference genome harbors minor alleles in over 2 million positions [84], which can severely impact downstream analysis [89]. For example, this study reported individual genomes will have ~30% false-positive and ~8% false-negative SNV calls due to these minor allele positions in the hg19 reference genome, with potential disease implications.

While the human reference genome is not a static entity, and new and improved versions are released regularly, further improvements to the reference genome are needed to optimize its

utility. For example, the current reference genome is based on a limited number of individuals, and therefore does not capture the full genomic diversity of the human population. As a result, *reference bias* occurs; samples that differ significantly from the reference genome from the reference genome may not map correctly, which in turn leads to incorrect variant calls [90]. One proposed solution to this problem involves the creation of population-specific reference genomes [91], however, the genetic ethnicity of individuals is not always easily determined. Therefore, we need a reference genome format that is able to capture the full genomic diversity of the human population. This requires a move away from the classic representation of the human reference genome as a single linear sequence, towards a graph-based model capable of capturing the full genetic diversity of humanity. Such a graph representation would allow for the inclusion of population-specific information, which could improve the accuracy of genomic analyses of ethnic minorities. This graph genome approach holds particular promise when it comes to structural variants, by allowing known SVs to be represented in the reference genome and thereby improving SV calling and characterisation [92]. The latest reference genome, hg38, supports such a graph-based representation by including the concept of alternative loci. However, the uptake has been slow due to the inertia of analysis tools to adapt to this shift in paradigm. But as the ecosystem of bioinformatics tools evolve to take full advantage of graph-based genomes, we will start to see improvements in variant calling accuracy.

The Informatics

Big and FAIR data

Data are being generated at an exponential rate, with estimates placing the storage capacity needed for human genomes alone as high as 40 exabytes in 2025 [93]. But storage space is only a small part of the story. The far greater challenge will lie in the data stewardship; the efficient handling of this data in a way that it can be easily searched, accessed, and shared within the worldwide scientific community. Initiatives that champion the FAIR data principles [94] are aimed at providing the necessary best practice guidelines needed for efficient data management and analyses in the genomic big data era. Beyond the technical challenges posed by this data explosion lie the even more complex and often legal questions of data privacy and security. These efforts must occur at different levels of organisation; each institute should provide guidelines to standardize data management within its various departments. National-level initiatives may then focus on interoperability between the various systems in use, which in turn can be used by international efforts to provide global data harmonization solutions. Many such international initiatives exist, such as the Global Alliance for Genomics and Health (GA4GH) [95], ELIXIR [96] and CINECA [97], which are currently coordinating their efforts to provide interoperability layers on top of existing data management

solutions. When generally adopted, these solutions will promote interoperability of data, and allow for a federated data analysis model.

Sustainability of bioinformatics tools

Bioinformatics tools often have a limited lifespan, due to a variety of reasons. And as the work in this thesis demonstrates, even with a strong focus on -and passion for- creating FAIR and accessible bioinformatics content, ensuring the long-term sustainability and availability of bioinformatics outputs remains a real challenge. In some cases the tools simply become obsolete, as the field of bioinformatics moves on (e.g. the Complete Genomics tools). In other cases, tool maintenance is abandoned in favor of a novel solution (e.g. iReport-like functionality has now been adopted and adapted into Galaxy itself). In yet other cases, the platforms originally used for sharing of the tools have since been disbanded (e.g. the CTMM-TraIT Tool Shed), making the tools harder to find (a shift towards viewing scientific publications as *living documents* that allow for post-publication updates could remedy this). However, one of the most important factors in the limited longevity of bioinformatics outputs, is the fact that the majority of such tools are developed in an academic setting. Academia often has an almost singular focus on journal publications as intellectual outputs, and the time and effort required for creating and maintaining high-quality bioinformatics software is often grossly underestimated and undervalued. Writing high-quality code takes time, whereas a *quick-and-dirty* solution can often suffice for the purposes of a publication. As a result, this often leads to bioinformatics tools that are difficult to re-use by others, and are all but abandoned after publication. If our goal is to improve the sustainability of bioinformatics applications, we need a collective shift in the academic mindset towards valuing high-quality code, its maintenance, and the community building around it as much as we value publications.

Galaxy

Galaxy has established itself as a user-friendly data analysis platform for researchers in the global scientific community. Going forward, enhancements to Galaxy could expand its utility outside of research and into the clinic. A simplified GUI and integrated result reporting are emerging as new areas of focus within the Galaxy community. Furthermore, to increase its appeal among bioinformaticians, further steps may be taken to improve interoperability with other workflow management systems such as Nextflow [98], SnakeMake [99] and the many other existing workflow management systems [100]. Initiatives such as CWL (Common Workflow Language) [101] exist to improve interoperability between these existing systems. Galaxy has been working towards supporting import and export of CWL-specified workflows, and if such a workflow data standard becomes widely accepted it could allow for workflows to be shared between different workflow

systems, reducing duplicated efforts within the bioinformatics community.

Training

As bioinformatics and data analysis skills are increasingly becoming prerequisites for scientific research, scalable and flexible educational resources are required for the training of researchers and clinicians dealing with these datasets. Delivery of training in remote or economically underprivileged areas poses additional challenges; such communities often lack experienced local instructors as well as the technical infrastructure required for large data analysis. The Galaxy platform provides a step towards alleviating the latter, by allowing its users to access powerful computer resources with nothing more than a web browser. One way to address the lack of access to experienced instructors in geographically or economically isolated areas is the delivery of hybrid trainings. In such an approach, instructors do not travel, but teach in front of a camera, and a live feed is streamed to one or more satellite locations. Each of these locations consists of a classroom with students and local supporting instructors who can communicate with the primary instructors. In the Gallantries project [102] we recently prototyped such an approach for delivery of Galaxy workshops. With both Galaxy itself and the associated training materials being freely available on the internet, this approach greatly reduced the effort and cost of workshop organisation. Instructors could be enlisted from various geographical locations without the funds and time usually required for travel. Within the Gallantries pilot project we tested this approach through 2 workshops with 5 satellite locations, and instructors residing in 4 different locations. In the coming years we will continue developing and scaling up this hybrid training approach for Galaxy. Furthermore, this project was a collaboration between the Galaxy and Carpentries [103] communities, optimally utilizing the combined knowledge and experience these two communities represent, and fostering closer ties between distinct but complementary scientific training communities.

Bioinformatics in the Clinic

As sequencing becomes faster and more affordable, and bioinformatics pipelines become more accessible, NGS pipelines are becoming increasingly valuable for direct clinical applications. Commercial solutions are often expensive, opaque, and hard to customize to the clinics specific experimental setup and needs. Here Galaxy represents a viable alternative. Galaxy is designed primarily for research applications, but in order to make Galaxy more suitable for use in the clinic, several enhancements can be considered. Firstly, where a high degree of flexibility is desirable in a research context, for clinical application the pipelines should be fixed, immutable. Therefore, a second, workflow-based user view to Galaxy is needed. In this user perspective, clinicians would

be presented with a simplified interface, containing only their personal set of preconfigured workflows. These workflows cannot be changed by the user, and only those parameters explicitly exposed by the workflow developer can be specified by the end-user. This will decrease the risk of accidental mistakes by the clinicians running the workflow, and will increase confidence in the system. Secondly, enhanced reporting capabilities should be tightly integrated with the Galaxy core framework, to replace tools such as iReport. This will allow for advanced features within the report, as well as increase the sustainability of the reporting framework. We have discussed both these points with the Galaxy core team, and they are part of the current Galaxy roadmap.

Furthermore, training is a vital component in bringing Galaxy to the clinic. This training should focus on familiarizing clinicians with the Galaxy system and interface, as well as the operational details of their workflows, and interpretation of the results. Bioinformatics pipelines are often perceived as a *black box* by non-expert users, but a basic understanding of the inner-workings of the analysis pipelines are often crucial for optimal interpretation of results. Training is the best way to shine a light in this black box, and illuminate bioinformatics.

Open Science

With new data being generated at exponential rates, the amount of new (biological) knowledge obtained from this data will also increase. Ideally, science should be a collective pursuit, and the sadly still-pervasive attitudes inhibiting the sharing of data for often irrational fears of being scooped should be replaced with attitudes geared toward complete transparency and adherence to open science principles [104]. This will require changes in attitudes not only in individual researchers, but also academia as a whole, with a vital role for scientific journals. This shift is already underway, as evidenced by the observation that academic journals increasingly are making submission of data to public databases a prerequisite for publication. Not only will this increase the reproducibility of results –one of the cornerstones of scientific research– but it will also increase the accuracy of results by allowing for more thorough peer review. Furthermore, it will increase the rate of knowledge acquisition by allowing for the integration of data from research institutes around the world. Similarly, analysis software and tools also greatly benefit from this open science attitude by enabling the entire global community to serve as potential source code reviewers and contributors, which will improve the code quality in ways a single developer in one lab can never hope to compete with.

"No man is an island, entire of itself; every man is a piece of the continent, a part of the main." - John Donne

Bibliography

[0] L. D. Stein, "Bioinformatics: alive and kicking," *Genome biology*, vol. 9, no. 12, pp. 1–2, 2008.

[1] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent, and A. Nekrutenko, "Galaxy: a platform for interactive large-scale genome analysis," *Genome Res*, vol. 15, no. 10, pp. 1451–1455, 2005.

[2] D. Blankenberg, G. Von Kuster, N. Coraor, G. Ananda, R. Lazarus, M. Mangan, A. Nekrutenko, and J. Taylor, "Galaxy: A web-based genome analysis tool for experimentalists," 2010.

[3] E. Afgan, D. Baker, M. Van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, *et al.*, "The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update," *Nucleic acids research*, vol. 44, no. W1, pp. W3–W10, 2016.

[4] "Clcbio genomics workbench." https://www.clcbio.com.

[5] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li, "Taverna: a tool for the composition and enactment of bioinformatics workflows," *Bioinformatics*, vol. 20, pp. 3045–3054, jun 2004.

[6] "Onlinehpc." https://onlinehpc.com.

[7] "Anduril workflow platform." https://www.anduril.org.

[8] K. J. van der Velde, F. Imhann, B. Charbon, C. Pang, D. van Enckevort, M. Slofstra, R. Barbieri, R. Alberts, D. Hendriksen, F. Kelpin, M. de Haan, T. de Boer, S. Haakma, C. Stroomberg, S. Scholtens, G.-J. van de Geijn, E. A. M. Festen, R. K. Weersma, and M. A. Swertz, "MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians," *Bioinformatics*, vol. 35, pp. 1076–1078, Aug. 2018.

[9] "Galaxy tool shed." https://toolshed.g2.bx.psu.edu/.

[10] "List of galaxy-related publications powered by zotero." https://www.zotero.org/groups/1732893/galaxy.

[11] T. Klingström, R. Hernández-de Diego, T. Collard, and E. Bongcam-Rudloff, "Galaksio, a user friendly workflow-centric front end for galaxy," *EMBnet. journal*, vol. 23, p. 897, 2017.

[12] T. C. Matthews, F. R. Bristow, E. J. Griffiths, A. Petkau, J. Adam, D. Dooley, P. Kruczkiewicz, J. Curatcha, J. Cabral, D. Fornika, *et al.*, "The integrated rapid infectious disease analysis (irida) platform," *bioRxiv*, p. 381830, 2018.

[13] F. Lemoine, D. Correia, V. Lefort, O. Doppelt-Azeroual, F. Mareuil, S. Cohen-Boulakia, and O. Gascuel, "Ngphylogeny. fr: new generation phylogenetic services for non-specialists," *Nucleic acids research*, vol. 47, no. W1, pp. W260–W265, 2019.

[14] K. Wolstencroft, S. Owen, O. Krebs, Q. Nguyen, N. J. Stanford, M. Golebiewski, A. Weidemann, M. Bittkowski, L. An, D. Shockley, J. L. Snoep, W. Müller, and C. Goble, "Seek: a systems biology data and model management platform," *BMC Systems Biology*, vol. 9, p. 33, Jul 2015.

[15] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra, "Circos: An information aesthetic for comparative genomics," *Genome Res*, vol. 19, no. 9, pp. 1639–1645, 2009.

[16] F. Paas, A. Renkl, and J. Sweller, "Cognitive load theory and instructional design: Recent developments," *Educational psychologist*, vol. 38, no. 1, pp. 1–4, 2003.

[17] "Google charts." https://galaxyproject.org/teach/gtn/.

[18] B. A. Grüning, J. Fallmann, D. Yusuf, S. Will, A. Erxleben, F. Eggenhofer, T. Houwaart, B. Batut, P. Videm, A. Bagnacani, *et al.*, "The rna workbench: best practices for rna and high-throughput sequencing bioinformatics in galaxy," *Nucleic acids research*, vol. 45, no. W1, pp. W560–W566, 2017.

[19] C. Blank, C. Easterly, B. Gruening, J. Johnson, C. Kolmeder, P. Kumar, D. May, S. Mehta, B. Mesuere, Z. Brown, *et al.*, "Disseminating metaproteomic informatics capabilities and knowledge using the galaxy-p framework," *Proteomes*, vol. 6, no. 1, p. 7, 2018.

[20] B. Batut, K. Gravouil, C. Defois, S. Hiltemann, J.-F. Brugère, E. Peyretaillade, and P. Peyret, "Asaim: a galaxy-based framework to analyze raw shotgun data from microbiota," *bioRxiv*, p. 183970, 2017.

[21] S. D. Hiltemann, S. A. Boers, P. J. van der Spek, R. Jansen, J. P. Hays, and A. P. Stubbs, "Galaxy mothur toolset (gmt): a user-friendly application for 16s rrna gene sequencing analysis using mothur," *GigaScience*, vol. 8, no. 2, p. giy166, 2018.

[22] "Jekyll: simple, blog-aware static sites." https://jekyllrb.com/.

[23] "Markdown." http://daringfireball.net/projects/markdown/.

[24] "Github: a web-based hosting service for version control using git." https://github.com/.

[25] "Travis ci - test and deploy your code with confidence." https://travis-ci.org/.

[26] T. N. Cuykendall, M. A. Rubin, and E. Khurana, "Non-coding genetic variation in cancer," *Current opinion in systems biology*, vol. 1, pp. 9–15, 2017.

[27] E. Khurana, Y. Fu, D. Chakravarty, F. Demichelis, M. A. Rubin, and M. Gerstein, "Role of non-coding sequence variants in cancer," *Nature Reviews Genetics*, vol. 17, no. 2, p. 93, 2016.

[28] P. C. Nowell and D. A. Hungerford, "Chromosome studies on normal and leukemic human leukocytes," *Journal of the National Cancer Institute*, vol. 25, no. 1, pp. 85–109, 1960.

[29] P. C. Nowell and D. A. Hungerford, "Chromosome studies in human leukemia. ii. chronic granulocytic leukemia," *Journal of the National Cancer Institute*, vol. 27, no. 5, pp. 1013–1035, 1961.

[30] B. J. Druker, C. L. Sawyers, H. Kantarjian, D. J. Resta, S. F. Reese, J. M. Ford, R. Capdeville, and M. Talpaz, "Activity of a specific inhibitor of the bcr-abl tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the philadelphia chromosome," *New England Journal of Medicine*, vol. 344, no. 14, pp. 1038–1042, 2001.

[31] B. J. Druker, M. Talpaz, D. J. Resta, B. Peng, E. Buchdunger, J. M. Ford, N. B. Lydon, H. Kantarjian, R. Capdeville, S. Ohno-Jones, *et al.*, "Efficacy and safety of a specific inhibitor of the bcr-abl tyrosine kinase in chronic myeloid leukemia," *N Engl j Med*, vol. 2001, no. 344, pp. 1031–1037, 2001.

[32] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X.-W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, *et al.*, "Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer," *science*, vol. 310, no. 5748, pp. 644–648, 2005.

[33] B. S. Taylor, N. Schultz, H. Hieronymus, A. Gopalan, Y. Xiao, B. S. Carver, V. K. Arora, P. Kaushik, E. Cerami, B. Reva, *et al.*, "Integrative genomic profiling of human prostate cancer," *Cancer cell*, vol. 18, no. 1, pp. 11–22, 2010.

[34] M. A. Rubin, C. A. Maher, and A. M. Chinnaiyan, "Common gene rearrangements in prostate cancer," *Journal of Clinical Oncology*, vol. 29, no. 27, p. 3659, 2011.

[35] I. Cortés-Ciriano, J. J.-K. Lee, R. Xi, D. Jain, Y. L. Jung, L. Yang, D. Gordenin, L. J. Klimczak, C.-Z. Zhang, D. S. Pellman, *et al.*, "Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing," *Nature genetics*, vol. 52, no. 3, pp. 331–341, 2020.

[36] N. A. Willis, E. Rass, and R. Scully, "Deciphering the code of the cancer genome: Mechanisms of chromosome rearrangement," *Trends in Cancer*, vol. 1, pp. 217–230, Dec. 2015.

[37] S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, *et al.*, "Mutational processes molding the genomes of 21 breast cancers," *Cell*, vol. 149, no. 5, pp. 979–993, 2012.

[38] C. F. Davis, C. J. Ricketts, M. Wang, L. Yang, A. D. Cherniack, H. Shen, C. Buhay, H. Kang, S. C. Kim, C. C. Fahey, *et al.*, "The somatic genomic landscape of chromophobe renal cell carcinoma," *Cancer cell*, vol. 26, no. 3, pp. 319–330, 2014.

[39] P. J. Stephens, C. D. Greenman, B. Fu, F. Yang, G. R. Bignell, L. J. Mudie, E. D. Pleasance, K. W. Lau, D. Beare, L. A. Stebbings, *et al.*, "Massive genomic rearrangement acquired in a single catastrophic event during cancer development," *cell*, vol. 144, no. 1, pp. 27–40, 2011.

[40] C. A. Maher and R. K. Wilson, "Chromothripsis and human disease: piecing together the shattering process," *Cell*, vol. 148, no. 1-2, pp. 29–32, 2012.

[41] M. M. Shen, "Chromoplexy: a new category of complex rearrangements in the cancer genome," *Cancer cell*, vol. 23, no. 5, pp. 567–569, 2013.

[42] C. Wu, A. W. Wyatt, A. McPherson, D. Lin, B. J. McConeghy, F. Mo, R. Shukin, A. V. Lapuk, S. J. M. Jones, Y. Zhao, *et al.*, "Poly-gene fusion transcripts and chromothripsis in prostate cancer," *Genes, Chromosomes and Cancer*, vol. 51, no. 12, pp. 1144–1153, 2012.

[43] S. C. Baca, D. Prandi, M. S. Lawrence, J. M. Mosquera, A. Romanel, Y. Drier, K. Park, N. Kitabayashi, T. Y. MacDonald, M. Ghandi, E. V. Allen, G. V. Kryukov, A. Sboner, J.-P. Theurillat, T. D. Soong, E. Nickerson, D. Auclair, A. Tewari, H. Beltran, R. C. Onofrio, G. Boysen, C. Guiducci, C. E. Barbieri, K. Cibulskis, A. Sivachenko, S. L. Carter, G. Saksena, D. Voet, A. H. Ramos, W. Winckler, M. Cipicchio, K. Ardlie, P. W. Kantoff, M. F. Berger, S. B. Gabriel, T. R. Golub, M. Meyerson, E. S. Lander, O. Elemento, G. Getz, F. Demichelis, M. A. Rubin, and L. A. Garraway, "Punctuated evolution of prostate cancer genomes," *Cell*, vol. 153, pp. 666–677, Apr. 2013.

[44] N. Voronina, J. K. Wong, D. Hübschmann, M. Hlevnjak, S. Uhrig, C. E. Heilig, P. Horak, S. Kreutzfeldt, A. Mock, A. Stenzinger, *et al.*, "The landscape of chromothripsis across adult cancer types," *Nature communications*, vol. 11, no. 1, pp. 1–13, 2020.

[45] A. S. Koltsova, A. A. Pendina, O. A. Efimova, O. G. Chiryaeva, T. V. Kuznetzova, and V. S. Baranov, "On the complexity of mechanisms and consequences of chromothripsis: An update," *Frontiers in Genetics*, vol. 10, Apr. 2019.

[46] W. P. Kloosterman, J. Koster, and J. J. Molenaar, "Prevalence and clinical implications of chromothripsis in cancer genomes," *Current opinion in oncology*, vol. 26, no. 1, pp. 64–72, 2014.

[47] M. C. Fontana, G. Marconi, J. D. M. Feenstra, E. Fonzi, C. Papayannidis, A. G. L. di Rora, A. Padella, V. Solli, E. Franchini, E. Ottaviani, *et al.*, "Chromothripsis in acute myeloid leukemia: biological features and impact on survival," *Leukemia*, vol. 32, no. 7, pp. 1609–1620, 2018.

[48] D. Hirsch, R. Kemmerling, S. Davis, J. Camps, P. S. Meltzer, T. Ried, and T. Gaiser, "Chromothripsis and focal copy number alterations determine poor outcome in malignant melanoma," *Cancer Research*, vol. 73, pp. 1454–1460, Dec. 2012.

[49] F. Magrangeas, H. Avet-Loiseau, N. C. Munshi, and S. Minvielle, "Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients," *Blood*, vol. 118, no. 3, pp. 675–678, 2011.

[50] J. Molenaar, J. Koster, D. Zwijnenburg, P. van Sluis, L. Valentijn, I. van der Ploeg, M. Hamdi, J. van Nes, B. Westerman, J. van Arkel, M. Ebus, F. Haneveld, A. Lakeman, L. Schild, P. Molenaar, P. Stroeken, M. van Noesel, I. Ora, E. Santo, H. Caron, E. Westerhout, and R. Versteeg, "Sequencing of neuroblastoma identifies chromothripsis and defects in neuritogenesis genes," *Nature*, vol. 483, pp. 589–593, Februari 2012.

[51] "Gnuplot." http://www.gnuplot.info.

[52] L. Oesper, A. Ritz, S. J. Aerni, R. Drebin, and B. J. Raphael, "Reconstructing cancer genomes from paired-end sequencing data," *BMC bioinformatics*, vol. 13, no. 6, p. S10, 2012.

[53] M. Lei, D. Liang, Y. Yang, S. Mitsuhashi, K. Katoh, N. Miyake, M. C. Frith, L. Wu, and N. Matsumoto, "Long-read dna sequencing fully characterized chromothripsis in a patient with langer–giedion syndrome and cornelia de lange syndrome-4," *Journal of Human Genetics*, pp. 1–8, 2020.

[54] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, *et al.*, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, p. 214, 2013.

[55] . G. P. Consortium *et al.*, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[56] . G. P. Consortium *et al.*, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, p. 56, 2012.

[57] R. Drmanac, A. Sparks, M. Callow, A. Halpern, N. Burns, B. Kermani, P. Carnevali, I. Nazarenko, G. GB Nilsen, G. Yeung, F. Dahl, A. Fernandez, B. Staker, K. Pant, J. Baccash, A. Borcherding, A. Brownley, R. Cedeno, L. Chen, D. Chernikoff, A. Cheung, R. Chirita, B. Curson, J. Ebert, C. Hacker, R. Hartlage, B. Hauser, S. Huang, Y. Jiang, V. Karpinchyk, and et al., "Human genome sequencing using unchained base reads on self-assembling dna nanoarrays," *Science*, vol. 327, pp. 78–81, January 2010.

[58] "Streeklab haarlem – laboratorium voor medische microbiologie." https://streeklabhaarlem.nl/.

[59] C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, J. H. Badger, A. T. Chinwalla, H. H. Creasy, A. M. Earl, M. G. FitzGerald, R. S. Fulton, *et al.*, "Structure, function and diversity of the healthy human microbiome," *nature*, vol. 486, no. 7402, p. 207, 2012.

[60] S. A. Boers, J. P. Hays, and R. Jansen, "Micelle pcr reduces chimera formation in 16s rrna profiling of complex microbial dna mixtures," *Scientific reports*, vol. 5, p. 14181, 2015.

[61] S. A. Boers, J. P. Hays, and R. Jansen, "Novel micelle pcr-based method for accurate, sensitive and quantitative microbiota profiling," *Scientific reports*, vol. 7, p. 45536, 2017.

[62] L. B. Price, C. M. Liu, J. H. Melendez, Y. M. Frankel, D. Engelthaler, M. Aziz, J. Bowers, R. Rattray, J. Ravel, C. Kingsley, *et al.*, "Community analysis of chronic wound bacteria using 16s rrna gene-based pyrosequencing: impact of diabetes and antibiotics on chronic wound microbiota," *PLoS One*, vol. 4, no. 7, p. e6462, 2009.

[63] S. A. Boers, E. I. Prest, M. Taučer-Kapteijn, A. Knezev, P. G. Schaap, J. P. Hays, and R. Jansen, "Monitoring of microbial dynamics in a drinking water distribution system using the culture-free, user-friendly, mycrobiota platform," *Scientific reports*, vol. 8, 2018.

[64] S. A. Boers, L. Reijnen, B. L. Herpers, J. P. Hays, and R. Jansen, "Detection of bacterial dna in septic arthritis samples using the mycrobiota platform," *J Clin Rheumatol. https://doi. org/10.1097/RHU*, vol. 901, 2018.

[65] T. Adékambi, M. Drancourt, and D. Raoult, "The rpob gene as a tool for clinical microbiologists," *Trends in microbiology*, vol. 17, no. 1, pp. 37–45, 2009.

[66] S. Yamamoto and S. Harayama, "Pcr amplification and direct sequencing of gyrb genes with universal primers and their application to the detection and taxonomic analysis of pseudomonas putida strains.," *Appl. Environ. Microbiol.*, vol. 61, no. 3, pp. 1104–1109, 1995.

[67] C. L. Schoch, K. A. Seifert, S. Huhndorf, V. Robert, J. L. Spouge, C. A. Levesque, W. Chen, F. B. Consortium, *et al.*, "Nuclear ribosomal internal transcribed spacer (its) region as a universal dna barcode marker for fungi," *Proceedings of the National Academy of Sciences*, vol. 109, no. 16, pp. 6241–6246, 2012.

[68] Y. Lan, G. Rosen, and R. Hershberg, "Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains," *Microbiome*, vol. 4, no. 1, p. 18, 2016.

[69] A. J. Sabat, E. van Zanten, V. Akkerboom, G. Wisselink, K. van Slochteren, R. F. de Boer, R. Hendrix, A. W. Friedrich, J. W. Rossen, and A. M. M. Kooistra-Smid, "Targeted next-generation sequencing of the 16s-23s rrna region for culture-independent bacterial identification-increased discrimination of closely related species," *Scientific reports*, vol. 7, no. 1, p. 3434, 2017.

[70] D. Sune, H. Rydberg, Å. N. Augustinsson, L. Serrander, and M. B. Jungeström, "Optimization of 16s rRNA gene analysis for use in the diagnostic clinical microbiology service," *Journal of Microbiological Methods*, vol. 170, p. 105854, Mar. 2020.

[71] A. Akram, M. Maley, I. Gosbell, T. Nguyen, and R. Chavada, "Utility of 16s rRNA PCR performed on clinical specimens in patient management," *International Journal of Infectious Diseases*, vol. 57, pp. 144–149, Apr. 2017.

[72] P. D. Schloss, S. L. Westcott, T. Ryabin, J. Hall, M. Hartmann, E. Hollister, R. Lesniewski, B. Oakley, D. Parks, C. Robinson, J. Sahl, B. Stres, G. Thallinger, D. Van Horn, and C. Weber, "Introducing mothur: open-source, platformindependent, community-supported software for describing and comparing microbial communities," *2. Appl Environ Microbiol*, vol. 75, pp. 7537–7541, 2009.

[73] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Krona: interactive metagenomic visualization in a web browser," *Encyclopedia of Metagenomics: Genes, Genomes and Metagenomes: Basics, Methods, Databases and Tools*, pp. 339–346, 2015.

[74] H. M. Bik and P. Interactive, "Phinch: An interactive, exploratory data visualization framework for–omic datasets," *bioRxiv*, p. 009944, 2014.

[75] D. McDonald, J. C. Clemente, J. Kuczynski, J. R. Rideout, J. Stombaugh, D. Wendel, A. Wilke, S. Huse, J. Hufnagle, F. Meyer, *et al.*, "The biological observation matrix (biom) format or: how i learned to stop worrying and love the ome-ome," *GigaScience*, vol. 1, no. 1, p. 7, 2012.

[76] "Docker: Package software into standardized units for development, shipment and deployment." https://www.docker.com/.

[77] I. Toma, M. O. Siegel, J. Keiser, A. Yakovleva, A. Kim, L. Davenport, J. Devaney, E. P. Hoffman, R. Alsubail, K. A. Crandall, *et al.*, "Single-molecule long-read 16s sequencing to characterize the lung microbiome from mechanically ventilated patients with suspected pneumonia," *Journal of clinical microbiology*, vol. 52, no. 11, pp. 3913–3921, 2014.

[78] O. Franzén, J. Hu, X. Bao, S. H. Itzkowitz, I. Peter, and A. Bashir, "Improved otu-picking using long-read 16s rrna gene amplicon sequencing and generic hierarchical clustering," *Microbiome*, vol. 3, no. 1, p. 43, 2015.

[79] A. L. Norris, R. E. Workman, Y. Fan, J. R. Eshleman, and W. Timp, "Nanopore sequencing detects structural variants in cancer," *Cancer biology & therapy*, vol. 17, no. 3, pp. 246–253, 2016.

[80] M. Nattestad, S. Goodwin, K. Ng, T. Baslan, F. J. Sedlazeck, P. Rescheneder, T. Garvin, H. Fang, J. Gurtowski, E. Hutton, *et al.*, "Complex rearrangements and oncogene amplifications revealed by long-read dna and rna sequencing of a breast cancer cell line," *Genome research*, vol. 28, no. 8, pp. 1126–1135, 2018.

[81] F. Kraft and I. Kurth, "Long-read sequencing in human genetics," *medizinische genetik*, vol. 31, no. 2, pp. 198–204, 2019.

[82] N. E. Navin, "The first five years of single-cell cancer genomics and beyond," *Genome research*, vol. 25, no. 10, pp. 1499–1507, 2015.

[83] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.*, "The sequence of the human genome," *science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[84] . G. P. Consortium *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, p. 68, 2015.

[85] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, *et al.*, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, no. 7571, p. 75, 2015.

[86] Y. A. Barbitoff, I. V. Bezdvornykh, D. E. Polev, E. A. Serebryakova, A. S. Glotov, O. S. Glotov, and A. V. Predeus, "Catching hidden variation: systematic correction of reference minor allele annotation in clinical variant calling," *Genetics in Medicine*, vol. 20, no. 3, p. 360, 2018.

[87] M. Koko, M. O. Abdallah, M. Amin, and M. Ibrahim, "Challenges imposed by minor reference alleles on the identification and reporting of clinical variants from exome data," *BMC genomics*, vol. 19, no. 1, p. 46, 2018.

[88] A. Ferrarini, L. Xumerle, F. Griggio, M. Garonzi, C. Cantaloni, C. Centomo, S. M. Vargas, P. Descombes, J. Marquis, S. Collino, *et al.*, "The use of non-variant sites to improve the clinical assessment of whole-genome sequence data," *PloS one*, vol. 10, no. 7, p. e0132180, 2015.

[89] S. Karthikeyan, P. S. Bawa, and S. Srinivasan, "hg19k: addressing a significant lacuna in hg19-based variant calling," *Molecular genetics & genomic medicine*, vol. 5, no. 1, pp. 15–20, 2017.

[90] S. Ballouz, A. Dobin, and J. A. Gillis, "Is it time to change the reference genome?," *Genome biology*, vol. 20, no. 1, pp. 1–9, 2019.

[91] Y. S. Cho, H. Kim, H.-M. Kim, S. Jho, J. Jun, Y. J. Lee, K. S. Chae, C. G. Kim, S. Kim, A. Eriksson, *et al.*, "An ethnically relevant consensus korean reference genome is a step towards personal reference genomes," *Nature communications*, vol. 7, p. 13637, 2016.

[92] G. Rakocevic, V. Semenyuk, W.-P. Lee, J. Spencer, J. Browning, I. J. Johnson, V. Arsenijevic, J. Nadj, K. Ghose, M. C. Suciu, *et al.*, "Fast and accurate genomic analyses using genome graphs," tech. rep., Nature Publishing Group, 2019.

[93] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big data: Astronomical or genomical?," *PLOS Biology*, vol. 13, p. e1002195, jul 2015.

[94] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, p. 160018, 2016.

[95] "Global alliance for genomics and health." https://www.ga4gh.org.

[96] "Elixir: A distributed infrastructure for life-science information." https://elixir-europe.org.

[97] "Cineca: Common infrastructure for national cohorts in europe, canada, and africa." https://www.cineca-project.eu.

[98] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nature biotechnology*, vol. 35, no. 4, p. 316, 2017.

[99] J. Köster and S. Rahmann, "Snakemake—a scalable bioinformatics workflow engine," *Bioinformatics*, vol. 28, no. 19, pp. 2520–2522, 2012.

[100] "Existing workflow systems." s.apache.org/existing-workflow-systems.

[101] P. Amstutz, M. R. Crusoe, N. Tijanić, B. Chapman, J. Chilton, M. Heuer, A. Kartashov, D. Leehr, H. Ménager, M. Nedeljkovich, *et al.*, "Common workflow language, v1. 0," 2016.

[102] S. Hiltemann, H. Rasche, B. Batut, M. Kuzak, and F. Psomopoulos, "The gallantries project: When galaxy meets carpentries to develop and deliver open training in life sciences," 2018.

[103] G. Wilson, "Software carpentry: getting scientists to write better code by making them more productive," *Computing in Science & Engineering*, vol. 8, no. 6, pp. 66–69, 2006.

[104] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, *et al.*, "Promoting an open research culture," *Science*, vol. 348, no. 6242, pp. 1422–1425, 2015.

A

Summary

DNA is often referred to as the *source code of life*; it encodes the proteins that control the functioning of our cells and plays a huge role in our health. The publication of the human reference genome in 2003, combined with sustained technological advances in genome sequencing ever since, have transformed the field biomedical research, and have led to an explosion of the amount of data being generated. However, scientists typically are not trained in the skills required to manage and analyse these large datasets. Furthermore, bioinformatics tools and workflows tend to be very complex, and often require programming skills to run. As a result, researchers often rely on bioinformaticians to perform the data analyses for them. This skills gap can lead to the bioinformatics data analysis feeling like a mystical *black box* to the researchers and clinicians tasked with interpreting the data analysis results. However, a basic understanding of the tools and processes that make up the analysis pipelines is often crucial for an accurate understanding of the analysis results. In this thesis, we aim to shine a light into this black box; to de-mystify bioinformatics, and increase the accessibility of tools and workflows in order to empower researchers and clinician to run their own data analyses. To this end, we first developed the required technical framework, both for the data analysis itself, and for the training of researchers and clinicians in the use of bioinformatics tools and workflows. We then proceeded to applied this framework and the Open Science methodology to a set of use cases, in prostate cancer analysis and microbiota profiling.

Chapter 0 gives a short introduction to genomics, including a brief history of genome sequencing. It also describes the bioinformatics challenges faced in the analysis of these often large and complex datasets, and provides a set of best-practice guidelines to facilitate accessibility, reproducibility and interoperability of bioinformatics processes. Finally, it provides a

short background for each of the use cases, including fusion genes in prostate cancer, and microbiota profiling.

In order to make bioinformatics analyses more accessible to the domain experts, some technical groundwork is required. This is provided in **Chapter 1**. Here we introduce the Galaxy platform, a web-based bioinformatics workflow engine that enables scientist to analyse large datasets with nothing more than a web browser. Once data has been analysed, it must be presented in a comprehensible way to the researchers and clinicians capable of interpreting the results. Efficient visualisation of data is crucial here. Therefore, as part of this chapter, we have integrated the Circos tool into the Galaxy framework. This tool allows us to plot genome-wide data in a single circular plot. By using Galaxy, this tool becomes accessible to non-bioinformaticians as well. Finally, we developed iReport, a tool within the Galaxy framework that allows the creation of custom web-reports for results reporting. This tool can combine results from any number of tools, and may be tailored to fit the user's need. Together, these components form the basis for bringing bioinformatics analysis to biomedical researchers and clinicians.

With the technical framework in place, the next important step is the training of the researchers and clinicians in the use of this platform, as well as in the relevant computational an informatics concepts that may impact interpretation of analysis results. To this end, in **Chapter 2**, we developed the Galaxy training repository, in close collaboration with the University of Freiburg. This project created a central repository for Galaxy-based training materials. It is a inherently community-driven project, and we put great effort into facilitating contributions from researchers and teachers around the globe.

Next, we applied this process for accessible bioinformatics to a number of use cases. In **Chapter 3** we examined fusion genes in a prostate cancer cell-line. First, we developed iFUSE, a web-based application for the visualisation and exploration of potential fusion genes. This application was used to identify a large number of fusion genes in the VCaP cell-line, and prioritize them on basis of confidence and impact for further confirmation. Furthermore, we determined that one of the arms of chromosome 5 had undergone chromothripsis. This is a shattering of the genome in a single catastrophic event, and the subsequent imprecise stitching back together of the chromosomes by the cell's repair mechanisms. Using the Circos tool we were able to visualize this very effectively. A second use case within the prostate cancer domain is described in **Chapter 4**. When sequencing tumour samples, typically a sample from healthy tissue of the same patient is also sequenced, in order to distinguish the cancer-specific (somatic) variants from the germline variants. However, such an associated normal sample is not always available. For these cases, we examined the feasibility of using a *virtual normal* instead. This is a set of healthy genomes from healthy, unrelated, genetically diverse individuals. To this end, we first integrated a suite tools for variant analysis and visualisation into Galaxy, and combined them into a workflow.

In addition to these two research-oriented use cases, **Chapter 5** describes a clinical use case. Here we developed the MYcrobiota platform for diagnostic microbiota profiling using 16S sequencing. This pipeline was developed for use by Streeklab Haarlem as an augmentation to their clinical diagnostic practices when traditional methods do not provide a clear answer. First, we integrated the full mothur suite of 125+ tools into Galaxy, and then combined them into a workflow. In order to accommodate the specific experimental setup employed by the Streeklab, we additionally developed a set of auxiliary Galaxy tools to supplement the standard analysis pipeline. These tools, as well as several preexisting Galaxy tools were integrated into an end-to-end workflow, with an iReport at the end for results reporting. This workflow was carefully tested and validated in collaboration with the Streeklab. We also developed Docker images with the full Galaxy setup so that it can be run in-house by the Streeklab.

For each of these use cases, we adhered to the bioinformatics best-practices and Open Science principles outlined in the introduction, and all the code is freely available in GitHub. Training materials have also been developed in order to aid not only our own researchers and clinicians in the use of the tools and workflows we developed, but also others around the world. All these training materials have deposited in the Galaxy training repository described in Chapter 2.

B

Samenvatting

DNA word vaak beschouwd als de *broncode van het leven*; het encodeert de eiwitten die onze celprocessen beïnvloeden, en speelt een belangrijke rol in onze gezondheid. De publicatie van het humane referentie genoom in 2003, in combinatie met continue technologische vooruitgangen sindsdien, hebben het veld van biomedisch onderzoek volkomen getransformeerd, en hebben geresulteerd in een explosie van de hoeveelheid data die gegenereerd wordt in de biomedische wetenschappen.

Wetenschappers worden echter meestal niet opgeleid in de benodigde vaardigheden om efficient om te gaan met deze grote hoeveelheden complexe datasets en ze te analyseren. Verder zijn bioinformatica tools en pijplijnen vaak erg complex, en zijn programmeervaardigheden meestal benodigd om ze te kunnen gebruiken. Hierdoor onstaat de situatie waarin onderzoekers vaak afhankelijk zijn van gespecialiseerde bioinformatici om hun analyses uit te voeren. Door deze vaardigheidskloof kan bioinformatica vaak aanvoelen als een *black box* voor de onderzoekers en clinici die de resultaten van deze analyse moeten interpreteren. Desalniettemin is een basiskennis van de achterliggende computationele concepten vaak essentieel voor de accurate interpretatie van de resultaten. In dit proefschrift trachtten wij de bioinformatische *black box* te illumineren, en de tools en pijplijnen toegankelijk te maken voor onderzoekers en clinici, zodat ze hun eigen data analyses weer uit kunnen voeren, zonder afhankelijk te zijn van tussenkomst van een bioinformaticus.

Hiervoor hebben we eerst de technische grondslag ontwikkeld, zowel voor het uitvoeren van de data analyses, alsmede het opleiden van onderzoekers en clinici in het gebruik ervan. Vervolgens hebben we dit technische framework toegepast via een set van wetenschappelijke casussen, in het gebied van prostaatkanker en micribiota analyse.

Hoofdstuk 0 geeft een korte algemene introductie tot genomics, waaronder een korte geschiedenis van sequencing technologieen. Hierbij worden ook de uitdagingen beschreven die we tegenkomen in de bioinformatica bij het analyseren van de resulterende grote en complexe datasets, en geven we een set richtlijnen voor het faciliteren van toegankelijke, reproduceerbare en interoperabele data analyse. Tot slot wordt hier een korte achtergrond gegeven voor ieder van de casussen.

Hoofdstuk 1 beschrijft de technische grondslag die we ontwikkeld hebben om de bioinformatica meer toegankelijk te maken voor de wetenschappelijke experts die de data genereren en de resultaten zullen interpreteren. Hier introduceren we allereerst Galaxy, een gebruiksvriendelijk data analyse platform waarmee wetenschappers makkelijk hun datasets kunnen analyseren met enkel een web browser. Na het analyseren van de data moeten de resultaten op een gestructureerde en overzichtelijke manier gepresenteerd worden aan de gebruiker. Visualizatie is hierbij cruciaal. Daarom hebben we als tweede deel van dit hoofdstuk de Circos tool in Galaxy geïntegreerd. Circos is een krachtige tool waarmee we genoom-wijde data efficient in een circulair plot kunnen weergeven. Normaalgesproken vergt deze tool gespecialiseerde kennis van bioinformatica, maar door deze binnen Galaxy beschikbaar te maken kan de tool ook gebruikt worden zonder deze technische kennis. Tot slot beschrijft dit hoofdstuk de iReport tool, waarmee gebruikers zelf web reports kunnen configureren binnen Galaxy om de resultaten van hun analyses weer te geven, wat precies afgestemd kan woorden op hun behoeften. Samen zorgen deze 3 componenten ervoor dat NGS analyses toegankelijk worden voor biomedische onderzoekers en clinici.

Met de technische grondslag gelegd, is de volgende belangrijke stap het opleiden van onderzoekers en clinici om hiermee om te gaan, zowel als de benodigde basiskennis op te doen over de achterliggende computationele concepten die een effect kunnen hebben op de interpretatie van de resultaten. In **Hoofdstuk 2** hebben we een de Galaxy Training Repository ontwikkeld, in nauwe samenwerking met de Universiteit van Freiburg. In dit project hebben we een centrale repository ontwikkeld voor het verzamelen en onderhouden van wetenschappelijke training materialen die gebruik maken van het Galaxy platform. Door de grootschalige aard van dit project, is het zodanig opgezet dat de gebruikersgemeenschap rond Galaxy het project samen draaiend kan houden, en op deze wijze gebruik gemaakt kan worden van de gecombineerde expertise van wetenschappers en opleiders over de hele wereld.

Vervolgens hebben we deze aanpak toegepast op een aantal casussen. In **Hoofdstuk 3** hebben we fusie genen in een prostaatkanker cellijn onderzocht. Hiervoor hebben we eerst een applicatie ontwikkeld (*iFUSE*) voor het visualiseren van structurele varianten en het identificeren van fusiegen kandidaten. Door gebruik van deze applicatie, in combinatie met de Circos tool, waren we in staat een groot aantal potentiële fusie genen te identificeren in de VCaP cellijn, en hebben we deze bevindingen in het lab kunnen bevestigen. Verder hebben we zo ook kunnen vaststellen dat de q arm van chromosoom 5 *chromothripsis* vertoonde, een fenomeen waarbij een deel van het genoom verbrijzeld wordt, waarna het genetische materiaal vervolgens op inaccurate wijze weer aan elkaar geplakt wordt door de reparatiemechanismen van de cel.

Een tweede casus binnen het prostaat kanker domein beschrijven we in **Hoofdstuk 4**. Wanneer we een tumor sequencen, wordt er normaal gesproken meestal ook gezond weefsel van dezelfde patiënt gesequenced. Hierdoor kunnen we de tumorcellen vergelijken met gezonde cellen, om zodoende de kanker-specifieke varianten te onderscheiden van de germline cellen. Echter, zo'n geassocieerd normaal sample is niet altijd beschikbaar. In zulke gevallen wilden we achterhalen of deze aanpak gesimuleerd kon worden door het gebruiken van een grote set van monsters afkomstig van gezonde, ongerelateerde, ethnisch diverse individuen. Hiervoor hebben we eerst een set analyse tools geïdentificeerd en ontwikkeld, en deze in Galaxy beschikbaar gemaakt, en ze gecombineerd tot een pijplijn.

Naast deze twee onderzoeks geörienteerde toepassingen, beschrijven we in **Hoofdstuk 5** een derde, clinisch geörienteerde casus. Hier hebben we het MYcrobiota platform ontwikkeld voor de profilering van microbiota via 16S rRNA sequencing

voor gebruik binnen de diagnostiek. Deze applicatie is ontwikkeld in samenwerking met het Streeklab Haarlem, als aanvulling op hun traditionele diagnostiek, wanneer de resultaten daarvan geen uitsluitsel geven.

Hiervoor hebben we eerst de complete mothur toolkit van 125+ tools in Galaxy geïntegreerd, en een workflow gemaakt die precies afgestemd is op de exacte experimentele setup van het Streeklab. Hiervoor moesten ook een aantal nieuwe tools geschreven worden als aanvulling op de standaard analyse procedure. Deze tools, in combinatie met reeds bestaande Galaxy tools, hebben we samengevoegd tot een pijplijn, met een iReport aan het eind voor het presenteren van de resultaten aan de clinici. Deze pijplijn is grondig getest en gevalideerd in samenwerking met de experts op het Streeklab, alvorens in gebruik genomen te worden. Om het mogelijk te maken om MYcrobiota binnenshuis te draaien op het Streeklab, hebben we Docker images ontwikkeld met alle benodigdheden om de applicatie gemakkelijk overal te kunnen draaien.

Voor al deze toepassing hebben we ons gehouden aan de richtlijnen die we hebben beschreven in de introductie, en alle code is compleet open en voor iedereen vrij te gebruiken. Ook hebben we voor iedere casus training materialen ontwikkeld, en ondergebracht in de Galaxy training repository beschreven in **Hoofdstuk 2**, waardoor ze niet alleen beschikbaar zijn voor onze eigen onderzoekers en clinici, maar voor de wereldwijde gebruikersgemeenschap.

C

List of publications

Publications 5, 6, 7, 9, 10, 15, 17, 19, 21, 20 are part of this thesis.

0. Astrid Heikema, Rick Jansen, Saskia Hiltemann, John Hays, Andrew Stubbs, WeFaceNano: a user-friendly pipeline for complete ONT sequence assembly and detection of antibiotic resistance in multi-plasmid bacterial isolates, *BMC Microbiology*, [in revision]
1. Beatriz Serrano-Solano, Anika Erxleben, Cristóbal Gallardo-Alba, Helena Rasche, **Saskia Hiltemann**, Melanie Föll, Matthias Fahrner, Mark J. Dunning, Marcel Schulz, Beáta Scholtz, Dave Clements, Anton Nekrutenko, Bérénice Batut, Björn Grüning, Fostering Accessible Online Education Using Galaxy as an e-learning Platform, *PLOS Computational Biology*, [accepted, March 2021]
2. Subina Mehta, Marie Crane, Emma Leith, Bérénice Batut, **Saskia Hiltemann**, Magnus Ø Arntzen, Benoit J. Kunath, Francesco Delogu, Ray Sajulga, Praveen Kumar, James E. Johnson, Timothy J. Griffin1, Pratik D. Jagtap ASaiM-MT: a validated and optimized ASaiM workflow for metatranscriptomics analysis within Galaxy framework, *F1000*, February 2021, doi: 10.12688/f1000research.28608.1
3. Willem de Koning, Milad Miladi, **Saskia Hiltemann**, Astrid Heikema, John P Hays, Stephan Flemming, Marius van den Beek, Dana A Mustafa, Rolf Backofen, Björn Grüning, Andrew P Stubbs NanoGalaxy: Nanopore long-read sequencing data analysis in Galaxy, *GigaScience*, October 2020, doi: 10.1093/gigascience/giaa105
4. Astrid Heikema, Deborah Horst-Kreft, Stefan Boers, Rick Jansen, **Saskia Hiltemann**, Willem de Koning, Robert

Kraaij, Maria de Ridder, Chantal van Houten, Louis Bont, Andrew Stubbs and John Hays. Comparison of Illumina versus Nanopore 16S rRNAGene Sequencing of the Human Nasal Microbiota, *Genes*, September 2020, doi: 10.3390/genes11091105

5. Helena Rasche, **Saskia Hiltemann**. Galactic Circos: User-friendly Circos plots within the Galaxy platform. *GigaScience*, June 2020, doi: 10.1093/gigascience/giaa065

6. **Saskia Hiltemann**, Stefan A Boers, Peter J van der Spek, Ruud Jansen, John P Hays, Andrew P Stubbs Galaxy mothur Toolset (GmT): a user-friendly application for 16S rRNA gene sequencing analysis using mothur. *GigaScience*, February 2019, doi: 10.1093/gigascience/giy166

7. Bérénice Batut, **Saskia Hiltemann**, Andrea Bagnacani, Dannon Baker, Vivek Bhardwaj, Clemens Blank, Anthony Bretaudeau, Loraine Guéguen, Martin Čech, John Chilton, Dave Clements, Olivia Doppelt-Azeroual, Anika Erxleben, Mallory Freeberg, Simon Gladman, Youri Hoogstrate, Hans-Rudolf Hotz, Torsten Houwaart, Pratik Jagtap, Delphine Lariviere, Gildas Le Corguillé, Thomas Manke, Fabien Mareuil, Fidel Ramírez, Devon Ryan, Florian Sigloch, Nicola Soranzo, Joachim Wolff, Pavankumar Videm, Markus Wolfien, Aisanjiang Wubuli, Dilmurat Yusuf, Rolf Backofen, Anton Nekrutenko, Björn Grüning Community-driven data analysis training for biology., June 2018, doi: 10.1016/j.cels.2018.05.012

8. Bérénice Batut, Kévin Gravouil, Clémence Defois, **Saskia Hiltemann**, Jean-François Brugère, Eric Peyretaillade, Pierre Peyret ASaiM: a Galaxy-based framework to analyze raw shotgun data from microbiota. *GigaScience*, June 2018, doi: 10.1093/gigascience/giy057

9. Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn A Grüning, Aysam Guerler, Jennifer Hillman-Jackson, **Saskia Hiltemann**, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, Daniel Blankenberg The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update , *Nucleic Acids Research*, May 2018, doi: 10.1093/nar/gky379

10. Stefan A. Boers, **Saskia D. Hiltemann**, Andrew P. Stubbs, Ruud Jansen and John P. Hays. Development and evaluation of a culture-free microbiota profiling platform (MYcrobiota) for clinical diagnostics, March 2018, doi: 10.1007/s10096-018-3220-z

11. Björn Grüning, Ryan Dale, Andreas Sjödin, Brad Chapman, Jillian Rowe, Christopher Tomkins-Tinch, Renan Valieris, Adam Caprez, Bérénice Batut, Mathias Haudgaard, Thomas Cokelaer, Kyle Beauchamp, Brent Pedersen, Youri Hoogstrate, Anthony Bretaudeau, Devon Ryan, Gildas Le Corguillé, Dilmurat Yusuf, Sebastian Luna-Valero, Rory Kirchner, Karel Brinda, Thomas Wollmann, Martin Raden, Simon van Heeringen, Nicola Soranzo, Lorena Pantano, Zachary Charlop-Powers, Per Unneberg, Matthias De Smet, Marcel Martin, Greg Von Kuster, Tiago Antao, Milad Miladi, Kevin Thornton, Christian LU Brueffer, Marius van der Beek, Daniel Maticzka, Clemens Blank, Sebastian Will, Kévin Gravouil, Joachim Wolff, Manuel Holtgrewe, Jörg Fallmann, Vitor Piro, Ilya Shlyakhter, Ayman Yousif, Philip Mabon, Xiao-Ou Zhang, Wei Shen, Jennifer Cabral, Cristel Thomas, Eric Enns, Joseph Brown, Jorrit Boekel, Mattias de Hollander, Jerome Kelleher, Nitesh Turaga, Julian de Ruiter, Dave Bouvier, Simon Gladman, Saket Choudhary, Nicholas Harding, Florian Eggenhofer, Arne Kratz, Zhuoqing Fang, Robert Kleinkauf, Henning Timm, Peter Cock, Enrico Seiler, Colin Brislawn, Hai Nguyen, Endre Bakken Stovner, Philip Ewels, Matt Chambers, James Johnson, Emil Hägglund, Simon Ye, Roman Valls Guimera, Elmar Pruesse, Augustine Dunn, Lance Parsons, Rob Patro, David Koppstein, Elena Grassi, Inken Wohlers, Alex Reynolds, MacIntosh Cornwell, Nicholas Stoler, Daniel Blankenberg, Guowei He, Marcel Bargull, Alexander Junge, Rick Farouni, Mallory Freeberg, Sourav Singh, Daniel Bogema, Fabio Cumbo, Liang-Bo Wang, David Larson, Matthew Workentine, Upendra Kumar Devisetty, Sacha Laurent, Pierrick Roger, Xavier Garnier, Rasmus Agren, Aziz Khan, John Eppley, Wei Li, Bianca Katharina Stöcker, Tobias Rausch, James Taylor, Patrick Wright,

Adam Taranto, Davide Chicco, Bengt Sennblad, Jasmijn Baaijens, Matthew Gopez, Nezar Abdennur, Iain Milne, Jens Preussner, Luca Pinello, Avi Srivastava, Aroon Chande, Philip Reiner Kensche, Yuri Pirola, Michael Knudsen, Ino de Bruijn, Kai Blin, Giorgio Gonnella, Oana Enache, Vivek Rai, Nicholas Waters, **Saskia Hiltemann**, Matthew Bendall, Christoph Stahl, Alistair Miles, Yannick Boursin, Yasset Perez-Riverol, Sebastian Schmeier, Erik Clarke, Kevin Arvai, Matthieu Jung, Tomás Di Domenico, Julien Seiler, Helena Rasche, Etienne Kornobis, Daniela Beisser, Sven Rahmann, Alexander Mikheyev, Camy Tran, Jordi Capellades, Christopher Schröder, Adrian Emanuel Salatino, Simon Dirmeier, Timothy Webster, Oleksandr Moskalenko, Stephen, Gordon and Köster, Johannes Bioconda: A sustainable and comprehensive software distribution for the life sciences., *Nature Methods*, July 2018, doi: 10.1038/s41592-018-0046-7

12. Chao Zhang, Jochem Bijlard, Christine Staiger, Serena Scollen4, David van Enckevort, Youri Hoogstrate, Alexander Senf, **Saskia Hiltemann**, Susanna Repo, Wibo Pipping, Mariska Bierkens, Stefan Payralbe, Bas Stringer, Jaap Heringa, Andrew Stubbs, Luiz Olavo Bonino Da Silva Santos, Jeroen Belien, Ward Weistra, Rita Azevedo, Kees van Bochove, Gerrit Meijer, Jan-Willem Boiten, Jordi Rambla, Remond Fijneman, J. Dylan Spalding, Sanne Abeln. Systematically linking tranSMART, Galaxy and EGA for reusing human translational research data. *F1000Research*, August 2017, doi: 10.12688/f1000research.12168.1.

13. Youri Hoogstrate, Chao Zhang, Alexander Senf, Jochem Bijlard, **Saskia Hiltemann**, David van Enckevort, Susanna Repo, Jaap Heringa, Guido Jenster, Remond J.A. Fijneman, Jan-Willem Boiten, Gerrit A. Meijer, Andrew Stubbs, Jordi Rambla, Dylan Spalding and Sanne Abeln. Integration of EGA secure data access into Galaxy. *F1000Research*. December 2016, doi: 10.12688/f1000research.10221.1.

14. Youri Hoogstrate, Rene Böttcher, **Saskia Hiltemann**, Peter van der Spek, Guido Jenster, Andrew Stubbs. FuMa: reporting overlap in RNA-seq detected fusion genes. *Bioinformatics*, April 2016, doi: 10.1093/bioinformatics/btv721.

15. **Saskia Hiltemann**, Guido Jenster, Jan Trapman, Peter van der Spek, Andrew Stubbs. Discriminating somatic and germline mutations in tumor DNA samples without matching normals. *Genome Research*, September 2015, doi: 10.1101/gr.183053.114.

16. Michael Moorhouse, David van Zessen, Hanna IJspeert, **Saskia Hiltemann**, Bas Horsman, Peter van der Spek, Mirjam van der Burg, Andrew Stubbs. ImmunoGlobulin galaxy (IGGalaxy) for simple determination and quantitation of immunoglobulin heavy chain rearrangements from NGS. *BMC Immunology*, December 2014, doi: 10.1186/s12865-014-0059-7.

17. **Saskia Hiltemann**, Youri Hoogstrate, Peter van der Spek, Guido Jenster, Andrew Stubbs. iReport: a generalised Galaxy solution for integrated experimental reporting. *GigaScience*, October 2014, doi: 10.1186/2047-217X-3-19.

18. Ines Teles Alves, Thomas Hartjes, Elizabeth McClellan, **Saskia Hiltemann**, Rene Böttcher, Dits N, Temanni MR, Janssen B, van Workum W, Peter van der Spek, Andrew Stubbs, de Klein A, Eussen B, Jan Trapman, Guido Jenster. Next-generation sequencing reveals novel rare fusion events with functional implication in prostate cancer. *Oncogene*, January 2015, doi: 10.1038/onc.2013.591.

19. **Saskia Hiltemann**, Hailiang Mei, Matthias de Hollander, Ivo Palli, Peter van der Spek, Guido Jenster, Andrew Stubbs. CGtag: complete genomics toolkit and annotation in a cloud-based Galaxy. *GigaScience*, January 2014, doi: 10.1186/2047-217X-3-1.

20. **Saskia Hiltemann**, Elizabeth McClellan, Jos van Nijnatten, Bas Horsman, Ivo Palli, Ines Teles Alves, Thomas Hartjes, Jan Trapman, Peter van der Spek, Guido Jenster, Andrew Stubbs. iFUSE: integrated fusion gene explorer. *Bioinformatics*, July 2013, doi: 10.1093/bioinformatics/btt252.

21. Ines Teles Alves, **Saskia Hiltemann**, Thomas Hartjes, Peter van der Spek, Andrew Stubbs, Jan Trapman, Guido Jenster. Gene fusions by chromothripsis of chromosome 5q in the VCaP prostate cancer cell line. *Human Genetics*, June 2013, doi: 10.1007/s00439-013-1308-1.

22. Andrew Stubbs, Elizabeth McClellan, Bas Horsman, **Saskia Hiltemann** , Ivo Palli, Stephan Nouwens, Anton Koning, Hoogland F, Reumers J, Daphne Heijsman, Sigrid Swagemakers, Kremer A, Meijerink J, Lambrechts D, Peter van der Spek. Huvariome: a web server resource of whole genome next-generation sequencing allelic frequencies to aid in pathological candidate gene selection. *Journal of Clinical Bioinformatics*, November 2012, doi: 10.1186/2043-9113-2-19.

D

Curriculum Vitae

Saskia Hiltemann was born on July 24, 1984, in Tilburg, The Netherlands. She completed high school (VWO) in 2002 at the Alfrink College in Zoetermeer. After a life-long fascination with the stars, she then started a bachelor in Physics and Astronomy at the University of Leiden. During her bachelor she took some programming classes and loved it so much that she switched to a Bachelor in Computer Science, also at Leiden University, which she completed in 2008. While she was passionate about computer science, she missed the science and math she had come to love when studying Astronomy, so she chose to pursue a Master's degree in Bioinformatics. There she focused heavily on the theoretical side of bioinformatics, with her Master's thesis on the topic of cellular automata and reaction-diffusion systems. Saskia also always had a love for teaching; all throughout her studies she worked as a tutor of high school students, helping them with math and science. After completing her Master's degree, she started working at the Erasmus Medical Center, on various project dedicated to building the infrastructure needed to analyse and interpret biological data, with a special focus on prostate cancer and later microbiota analysis. During her time at Erasmus MC she discovered a passion for the world of open-source software, and has contributed to many open source projects. She also co-founded the Galaxy Training Network, bringing together her love for teaching, science, and open source software.

E

Phd Portfolio

Name PhD student	Saskia Hiltemann
PhD period	2012-2016
Erasmus MC Department	Urology and Bioinformatics
Research School	MolMED
Promotor	Prof. Guido Jenster and Prof. Peter van der Spek
Supervisor	Prof. Guido Jenster and Dr. Andrew Stubbs

1. PhD training

General and Specific courses (ECTS)

2012	NCSB Tutorial: Statistics with R	VU
2012	NBIC Genomic Resequencing	Nijmegen
2012	EBI Roadshow	EMC
2012	A first Encounter with Next-generation Sequencing Data	EMC
2013	NBIC Advanced de novo Assembly	Wageningen
2014	SURFSara Grid computing MOOC	Online
2015	Molmed Basic and Translational Oncology	EMC

Seminar Series

2012-2016	NBIC programmers meeting	Utrecht
2012-2016	Bridge meetings	EMC
2012-2016	JNI meetings	EMC

2. Presentations

2012-2016	Urology Lab meetings	EMC	(t)
2012-2020	Bridge meetings	EMC	(t)
2012-2016	JNI meetings	EMC	(t)
2012-2018	NBIC/DTL programmers meetings	Utrecht	(t)

National and international conferences

2012-2015	TraIT annual meeting	Utrecht	(p,s)
2013,2014	NBIC Conference	Lunteren	(p)
2012	Unanswered Questions in Cancer Sequencing	Cambridge, UK	(p)
2012	Benelux Bioinformatics Conference	Nijmegen	(p)
2013	Galaxy Community Conference	Oslo, Norway	(p)
2014	Molecular Medicine Day	EMC	(p)
2014	2016 Daniel den Hoed Day	EMC	(p)
2014	Benelux Bioinformatics Conference	Luxemburg, LU	(p)
2014	Galaxy Community Conference	Baltimore, USA	(t,p,w,s)
2015	Galaxy Community Conference	Norwich, UK	(p,h,t,w,s)
2015	25th MGC symposium	Leiden	(t)
2016	Galaxy Community Conference	Indiana, USA	(p,h,w,o)
2017	Galaxy Community Conference	Montpellier, FR	(p,h,w)
2017	Health-RI Conference	Utrecht	(p)
2017	27th MGC symposium	Leiden	(p)
2018	GalaxyEU User Conference	Freiburg	(p,t)
2018	Galaxy Community Conference	Portland, USA	(o,w)
2018	Biohackathon	Paris, FR	(h)
2019	Galaxy Community Conference	Freiburg, DE	(o,w,h)
2019	Biohackathon	Paris, FR	(h)
2020	ABRF Conference	Palm Springs, USA	(w)
2020	Galaxy Community Conference	Virtual	(w,o,h)
2020	BioHackathon	Virtual	(h)
2021	EOSC-Life Remote Training series	Virtual	(t,s)
2021	Galaxy Community Conference	Virtual	(t,w,p,o,s)
2021	ISMB/ECCB WEB	Virtual	(t)

(p=poster, t=talk, s=software demo, w=workshop, h=hackathon, o=organizer)

3. Teaching - Highlighted activities

(co-) Supervising Master's theses
2012-2016 - Jos, Daniel, Jeroen, Rick, Willem

2013	Galaxy Toolshed	NBIC programmers meeting
2014	Introduction to Galaxy	TraIT Consortium
2014	RNAseq Data analysis	LUMC
2014	RNA-seq analysis	Galaxy Community Conference
2015	Introduction to Galaxy	EMC
2015	Introduction to Galaxy	BSc course Delft
2015	RNAseq Data analysis	LUMC
2015	RNA-seq analysis	Galaxy Community Conference
2015	Genomic Resequencing in Medical Diagnostics	EMC
2016	RNA-seq analysis	Galaxy Community Conference
2017	Elixir Galaxy Developers Workshop	Strassbourg, France
2017	Visualization of BIG DATA	Galaxy Community Conference
2017	Visualization development in Galaxy	Galaxy Community Conference
2017	Metagenomics: the full picture	Galaxy Community Conference
2017	BioSB RNA-Seq	LUMC
2017	Workshop FAIR data with myFAIR	Lygature Partners meeting
2018	Galaxy developer workshop	Melbourne, Australia
2018-present	Quarterly GTN Collaboration Fest lead	Virtual
2019	Gallantries Workshop (x3)	Distributed Training
2019	Metagenomics data analysis with Galaxy	Galaxy Community Conference
2019	Train the Galaxy Trainer	Galaxy Community Conference
2020	Galaxy administrator workshop	Barcelona, Spain
2020	Metatranscriptomics with Galaxy	ABRF, Palm Springs (USA)
2020	Train the Galaxy Trainer	Galaxy Community Conference
2020	Metatranscriptomics with Galaxy	Galaxy Community Conference
2020-present	Bioinf weekly git-together	EMC
2021	Galaxy administrator workshop	Virtual (also organizer)
2021	GTN Smorgasbord	Virtual (also organizer)
2021	Galaxy Resources for Trainers and Educators	Virtual
2021	Microbiobe metatranscriptomics analysis in Galaxy	Virtual ELIXIR-Belgium Workshop
2021	GCC Training Week	Galaxy Community Conference (also organi

4. Other highlighted activities

2016-present	Galaxy IUC	Committee member
2018-present	Galaxy Training Network	Founder & co-lead
2019-2023	CINECA Project	WP co-lead
2021-present	Galaxy Training & Outreach WG	Working group member
2019-present	Gallantries Project	Project Coordinator & WP lead

Grants obtained

2019	Mozilla mini-grant (7,5K)	Gallantries 1 year pilot project
2020	Erasmus Plus KA02 grant (400K)	Gallantries 3 year follow-up project

Acknowledgments

Team work makes the dream work. I have worked towards this thesis over the past 9 years at Erasmus MC, and in that time have had the distinct pleasure of working with many amazing people. A big thank you to everybody who was part of my journey, thanks to all my co-authors from around the world, and I would like to thank the following people in particular:

Dr Stubbs, Dear Andrew. Your guidance over the past 9 years shaped me into the scientist I am today. It is truly a pleasure working with you; thank you for your patience, easy-going demeanor, low-stress working environment, the many laughs, and for giving me the space to find my own path and building my confidence.

Prof. Jenster, dear Guido. Thank you for your guidance during my first years at EMC. You were never too busy to help when I had questions and always happy to explain the biology when I felt lost. Thanks for ending every email with a reminder to have fun, I needed it sometimes, but overall I succeeded and had a great deal of fun during my time at EMC.

Prof. van der Spek, dear Peter. I would like to thank you for your guidance during the final phase of my promotion, your useful feedback and suggestions, as well as the help in navigating the practicalities around my promotion during the pandemic are greatly appreciated.

Other committee members. Thanks for taking the time to carefully read my manuscript and providing insightful feedback that helped improve this thesis.

Coworkers. Thanks to everybody at the EMC for the years of fun, drinks, laughs, commiserations, rants, discussions, movie nights, dinners, even during a year of lockdown, it's been a blast!

Galaxy community. Thanks to everybody in the Galaxy community for making me feel welcome, and patiently helping me whenever I had questions. I learned lot from contributing to this project, and I've come to love the Galaxy community so much. A special thanks to Bérénice Batut, it's been a pleasure building the GTN and watching it grow over the years. Big shout out to Dave Clements for being the friendly face of an awesome community, you've been a great help and are always delightful to talk you.

Women in Science I feel lucky to work with so many amazing and strong women in science; Helena, Bérénice, Ines, Vera, Leslie, Subina, Beatriz, Wendi, Jen, Delphine, Anna, Anne, Maria, Elizabeth and many more, you are an inspiration and I've learned a bunch from all of you!

Family. Thanks to my Mom and Dad and two amazing sisters Miriam and Yvonne for always cheering me on, I love you all.

Helena. Last but not least, thanks to my amazing partner Helena for always being there for me, thanks for always reminding me that I can do this and having my back. Thanks for all the valuable discussions when I was writing this manuscript, and for co-authoring some of the papers in this thesis. Thanks also for helping me with the design of this thesis and your many opinions on fonts. Thanks for being my team mate, I love you.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, "Science Experiment 02", was created by Ben Schlitter and released under CC BY-NC-ND 3.0.