Summary

Bioinformatics is the interdisciplinary field that aims to develop and apply computational methods to tackle research questions in biology and genetics. This new field spans more than three decades but it has played a pivotal role in advancing the existing knowledge and practices in a range of biomedical and genetics domains. Today, there is a consensus that the solutions developed in this area have reached a long-awaited point where certain criteria regarding software quality, efficiency and availability are finally being met. In brief, these criteria concern the application of professional software development methods, the adaptation of user-friendly interfaces, the availability of easily accessible documentation, the effective use of web technologies, and finally, perhaps most important, the release of these solutions through open source licensing. Despite this highly significant milestone, the recent advances in mass genomic profiling technologies (i.e. DNA sequencing) have introduced an additional need that goes beyond the criteria that characterize the maturity level of any single tool. This additional need is integration. While the complexity, robustness and scalability of software methods in bioinformatics is steadily rising, we observe a concurrent increase of demand for solutions that glue these tools together. Integrating even seemingly unrelated tools has always been a central task of bioinformatics studies. Yet lately, we are experiencing an explosion of qualitative, general-purpose, community-developed and environment-agnostic tools that can boost the efficiency of analysis immensely and therefore the scientific value of existing bioinformatics tools in general. Without being exhaustive, we can place these tools in categories like simple scripts in modern interpreted languages, data visualization, tools for data annotation, validation and quality control, data management frameworks, management of jobs in High Performance Computational (HPC) environments, and tools that handle complete virtualized operating systems. Today, given the extreme volume and complexity of data in the life sciences, the majority of scientific tools in computational biology (e.g. for genotype imputation) can generate meaningful results only if they are placed in analysis pipelines that build on synergies with these general-purpose tools.

The main predicament for the work I have described in this thesis was to provide solutions for common bioinformatics problems in this new scientific terrain. In Chapter 1, I describe in detail how bioinformatics is currently entering a new era, after
establishing open source licenses as standard practice and with existing analysis tools reaching professional maturity. This new era demands co-operative solutions, an extrovert mentality, and reproducible experiments. I also present the main challenges, which are improving of IT skills and literacy in the life sciences, building co-operative computational infrastructures, and generating incentives that will encourage scientists to publish their source code, methods and data. Moreover, I describe the four main practical considerations that need to be addressed for making a bioinformatics component (i.e. tools, data) as useful as possible for modern research. These considerations are Documentation, Wrapping, Collaboration and Composition.

In Chapter 2, I present an overview of existing scientific workflow management systems in bioinformatics along with their pros and cons. I also present some good practice guidelines that can enhance the pipeline-ability of scientific software and data. I extend these guidelines for future workflow environments, in which I advocate openness, standardization, being able to embed arbitrary tools while also offering the ability to being embedded in other tools, encouraging user cooperation, support for HPCs and enabling virtualization. In chapter 2, I also present the foreseeable benefits of adopting these guidelines, the most important being increased reproducibility which could, in turn, bring personalized clinical genetics closer to reality.

In Chapter 3, I present a detailed computational pipeline for genotype imputation. This pipeline is essential in modern population genetics and phenotype-genotype association studies. I discuss several considerations, such as choice of existing software, choice of reference panel, tuning of parameters, quality control, assessment of results and visualization. This chapter also presents guidelines for the construction of a novel imputation reference set based on the Genome of the Netherlands (GoNL). This population-specific reference panel has been proved to increase significantly the imputation quality in Dutch cohorts and has helped to reveal additional genetic markers for known diseases. A priority for this chapter is to present in detail the command lines and computational requirements that can help even novice users to perform genotype imputation.

As a continuation to chapter 3, I present MOLGENIS-Impute in Chapter 4; this is an integrated genotype imputation pipeline based on the MOLGENIS-compute pipeline management system. It is a highly customizable, self-deployed imputation pipeline that requires absolutely no knowledge of the specifics of the underlying software. It is accompanied by additional tools for format conversion and quality control while it also manages the submission of computation tasks in a variety of HPC environments. It is intended to act as a one-stop solution for researchers who want to apply imputation as an intermediate step in their analysis.

In Chapter 5, I present PyPedia, which is a new concept in the area of scientific
computing. PyPedia utilizes the concept of wikis to offer a unified, inter-connected
development environment. Instead of building isolated solutions, PyPedia encourages
users to contribute either by creating new methods or by improving existing methods,
in the same fashion that wikis create qualitative content through crowdsourcing.
PyPedia users can contribute in a variety of ways, according to their expertise (i.e.
source, tests, documentation). All content is public and execution can take place in
HPC environments, local computers, online, or in a specially designed virtualized
environment (Docker). All the methods developed in chapters 3 and 4 are also available
in PyPedia.

In Chapter 6, I present a new pipeline, MutationInfo, for a problem that is pertinent
in clinical genetics, i.e. to efficiently locate the position of genetic variants that are
published in locus-specific databases or scientific reports. This task is crucial in efforts
that try to validate the existence of an already published variant (or one under
investigation) in a sequenced or genotyped sample. The pipeline combines 11 different
tools or databases to optimally perform this task. As with MOLGENIS-impute, it is
self-deployed and requires limited IT knowledge. MutationInfo is also operating as an
online web service.

Finally in Chapter 7, I present my concluding remarks on genotype imputation,
focusing on the main practical challenges and future prospects of this method. I also
show how clinical genetics can be brought closer to mainstream medical practice by
integrating existing bioinformatics components such as data, tools and workflows.