7 Discussion

7.1 Final notes in Imputation

7.1.1 Imputation as a benchmark workflow

In previous chapters I have demonstrated that genotype imputation (or simply ‘imputation’) is a vital part of modern genetic analysis. In the context of designing a workflow environment, imputation is also a valuable testbed for benchmarking. Although a complete, sharable genotype imputation pipeline contains many challenges [32], it is important to note that doing imputation is a relatively easy task. Setting up the required tools and configuration should not take more than a few days for an average IT-skilled person (although chunking, quality control, and submission to a High Performance Computer (HPC) could increase this time to more than a month...).

Thousands of published studies that include imputation as part of their analysis have probably not used a “ready-to-use” pipeline, but have setup and configured one for each different study. Given that setting up an imputation pipeline is rather a tedious and uninteresting task for most biology or medical researchers, I consider that collectively, a substantially amount of time (and funding) is being wasted on tasks like this. Arguably, a large part of bioinformatics should focus on automating uninteresting tasks, giving researchers the luxury to delve more into interesting subjects. So offering workflow environments and creating automatic pipelines could significantly aid scientific progress.

To my knowledge, there are currently two imputation environments similar to the one presented here. The first is EZIMPUTER [72], which is more of a set of guidelines and scripts rather than a ready-to-run imputation pipeline. The second is Michigan Imputation Server [9], which is a very well-designed web framework that offers “imputation as a service”. Users have the options to either upload their data to a dedicated server or download a virtualized operating system (in the form of a Docker Image), which contains pre-installed and pre-configured all-required software. The only limitation of this approach is that users do not have the option to submit the analysis to a custom HPC environment, for example, in a department’s cluster. So one question could be: Why were complete imputation pipelines not widely available before? The answer lies in Albert Einstein’s famous quotation: “Make things as simple as possible,
but not simpler”. Before attempting a simplification (or automation), we should be
confident that the underlying components are mature, stable and not subject to change.
We should also make sure that the offered abstraction makes sense to the community.
Over-simplification, over-packaging and over-automation have the serious hazard of
creating un-configurable and rigid environments with limited real value. The need to
“automate” imputation came right after the first data releases of the 1000 Genomes
Project (or else, 1000 GP) \[17\] in 2008. 1000 GP data were welcomed by the genetics
community as a quantitative alternative to the only existing imputation reference panel
at that time, namely, the HapMap project. The existence of two inherently different, but
equally important, imputation reference panels created the need to impute studies with
both of them and to compare the results. These comparisons gave great insights into
the quality of 1000 GP data and the merit of imputation in general. Thus, automation
emerged as a community need and part of the work in this thesis has tried to meet
this need. Of course, this automation is far from complete since we expect future
imputation pipelines to be more diverse and to include more exotic ideas. For example,
two-strategy imputation approaches (first with population-specific data and then with
1000 GP \[37\]), imputation of p-values from public GWA Studies \[38\], and imputation
of species for which there is no comprehensive reference panel (as yet) \[70\], \[81\].
To illustrate the complexity of setting up an imputation pipeline, I present table \ref{pipeline}.
In this table I list some of the most commonly used tools that researchers need for
such a setup. In total, there are hundreds of pages of documentation that someone has
to go through for setting up a pipeline correctly. Moreover, few tools are written in
programming languages that prioritize code readability and even fewer are available
in a consistent, browsable repository like GIT or Mercurial. The only tool written in
an easily readable language in GIT is Genotype Harmonizer \[12\]. Besides, the most
prominent tool in imputation, IMPUTE2, has not made its source code open. So being
able to setup a pipeline still requires a basic understanding of C, C++, Python, Java
and BASH scripting, and sufficient experience to comprehend technical IT manuals.
This table does not include other necessary tools for sideways analysis, like format
conversion (vcftools, PLINK, tabix) or downstream analysis (e.g. SNPTEST \[51\] and
meta-analysis methods \[11\]). So although most of the tools are of quite good quality
and efficiency, they are far from being easy-to-incorporate into pipelines.
A final note on automation is that the choice of imputation tool is actually of
secondary importance. I present a thorough discussion on this in the next section.

\footnote{\url{http://hgdownload.cse.ucsc.edu/downloads.html}}
\footnote{\url{http://genome.sph.umich.edu/wiki/CheckVCF.py}}
\footnote{\url{http://www.well.ox.ac.uk/~gav/qctool/#overview}}
# 7.1 Final notes in Imputation

<table>
<thead>
<tr>
<th>Tool</th>
<th>Steps</th>
<th>Open Source</th>
<th>GIT/Mercurial</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCSC LiftOver(^1)</td>
<td>Liftover</td>
<td>Yes (C)</td>
<td>No</td>
</tr>
<tr>
<td>Crossmap [85]</td>
<td>Liftover</td>
<td>Yes (C, Python)</td>
<td>No</td>
</tr>
<tr>
<td>GACT [75]</td>
<td>Liftover, QC</td>
<td>Web tool,</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes (BASH)</td>
<td></td>
</tr>
<tr>
<td>CheckVCF.py (^2)</td>
<td>QC</td>
<td>Yes (Python)</td>
<td>Yes</td>
</tr>
<tr>
<td>QCTOOL (^3)</td>
<td>QC</td>
<td>Yes (C++)</td>
<td>Yes</td>
</tr>
<tr>
<td>Genotype Harmonizer [12]</td>
<td>QC, Format Conversion</td>
<td>Yes (Java)</td>
<td>Yes</td>
</tr>
<tr>
<td>Beagle</td>
<td>Phasing [6],</td>
<td>Yes (Java)</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Imputation [5]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMPUTE2 [29]</td>
<td>Imputation</td>
<td>Closed source</td>
<td>No</td>
</tr>
<tr>
<td>Minimac2 [28]</td>
<td>Imputation</td>
<td>Yes (C)</td>
<td>Yes</td>
</tr>
<tr>
<td>Fish [84]</td>
<td>Imputation</td>
<td>Closed source</td>
<td>No</td>
</tr>
<tr>
<td>Shapeit [14]</td>
<td>Phasing</td>
<td>Closed source</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 7.1: The open source status of major tools in imputation.
Instead of looking for the ‘right’ imputation tool, a researcher should focus more on the correct tuning of the tool in the available HPC environment. The latest (January 2016) high-profile publication on imputation [5] focuses more on advances in computational resources and scaling options than on the increase in imputation quality. Other issues that also need to be dealt with in the future are multi-allelic and allosome (chromosomes X and Y) imputation.

7.1.2 Factors affecting imputation quality

As intuitively expected, and also demonstrated in numerous studies, imputation accuracy depends on three major factors. These are the quality of the study, the quality and size of the reference panel, and the relevance of the reference panel to the study population.

Regarding the quality of the study, Nelson et al. [58] demonstrated that array density (the number of SNPs that a platform contains) is not an important factor for imputing SNPs with a MAF higher than 0.05, as long as it exceeds 1 million markers. For SNPs with \(0.01 < \text{MAF} \leq 0.05\) density becomes irrelevant after 2.5 million markers. For more rare SNPs (\(\text{MAF} \leq 0.01\)), the imputation quality is linearly related to the density of the SNP array. Apart from the density of the platform, a researcher should also take into account the coverage that it offers for a particular population. Coverage is defined as the fraction of all SNPs in the genome that can be captured by the chip [44]. Different SNP arrays of the same size exhibit different degrees of coverage for European, Asian and African populations [23]. Consequently, some SNP arrays are better suited for imputation for specific populations [58]. For these comparisons, both coverage analysis [23] and density analysis [58] were performed with the 1000 GP serving as the imputation reference set.

Regarding the imputation reference set, today (July 2017) perhaps the largest known imputation reference panel for a single population (European) comprises 64,976 haplotypes and contains 39 million SNPs [19]. This reference panel has been assembled by the Haplotype Reference Consortium (HRC) \(^4\). In comparison, the 1000 Genomes Projects, in its last phase, contains 5,008 haplotypes and 88 million SNPs. The accurate imputation of very rare SNPs, even at the level of Minor Allele Frequency (MAF) of 0.1%, is practically possible with this reference panel. Before HRC, the largest imputation reference panel for European populations was the UK10K [30], which contained 7,562 haplotypes and 26 million variants. Despite being smaller in haplotype size, the 1000 Genomes Project contains more SNPs, since it includes samples from

\(^4\)http://www.haplotype-reference-consortium.org/
26 different populations. HRC also contains the UK10K, 1000 GP, and Genome of the Netherlands (GoNL) [4] reference sets, among others. Comparisons between the largest three reference panels (HRC, UK10K, and 1000 GP) showed that for SNPs with $MAF \geq 5\%$, all panels exhibit approximately the same imputation accuracy. This accuracy is measured as the squared correlation ($R^2$) between the imputed allele dosages and the real genotypes. Nevertheless, the accuracy gains become greater as the reference set becomes larger and the MAF smaller. To put this in perspective, the average imputation accuracy of a SNP with $MAF = 0.03\%$, imputed with HRC is the same as the average accuracy of a SNP with $MAF = 0.2\%$ imputed with 1000 GP. In other words, if in a study we filter out SNPs that have been imputed with an accuracy $R^2 < 0.8$ then with 1000 GP we would have to remove (on average) any SNPs with $MAF < 2\%$ whereas if the study is imputed with HRC, we would remove only SNPs with $MAF < 0.2\%$, a 10-fold difference. This shows the tremendous effect of the size of the imputation reference panel on rare alleles.

A vivid scientific discussion at the moment refers to the population composition of reference panels. Should we carefully pre-select the samples in a reference panel according to the population of interest in a GWAS or should we use universal reference panels? The authors of one the most prominent imputation tools, IMPUTE2 [27], advocate the latter approach since they incorporate methods that pre-select the most relevant haplotypes from the reference panel. However, the authors of minimac2 [20], another prominent tool, argue that, if possible, population selection should be performed although their tool is computationally efficient for very large reference panels. Universal reference panels are preferable for admixed populations. In this case IMPUTE2 can handle population admixture by default, whereas an alternative choice is to use MaCH-Admix [48]. A good practice that also reduces execution time is the pre-phasing of the study panel. IMPUTE2 is well coupled with SHAPEIT [14] for this purpose, whereas the MaCH-Minimac framework [45] offers the same functionality.

Regarding the quality of the reference panel, Deelen et al. [13] performed a comprehensive study that demonstrated the effect of a qualitatively crafted imputation reference panel targeting for a specific population. In this study, the reference panel was based on the Genome of the Netherlands (GoNL) [4], which contains 769 Dutch samples sequenced at an average depth of 14X. Apart from the relatively high depth, another comparative advantage of GoNL is that it is mainly composed from 231 family trios and 19 quartets. Since the phasing procedure incorporates information from one of the children [54], the resulting haplotypes are of very high quality. Results show an improved accuracy ($R^2$) of imputation for the Dutch samples: from 0.61 to 0.71 for rare variants ($0.05\% < MAF < 0.5\%)$. Roughly, the range of this improvement is comparable to that one observed when the size of the reference panel is increased.
from 1,000 to 5,000 haplotypes (Table 1 [20]). Yet the size of GoNL was 998 unrelated haplotypes, while the size of 1000 GP was 758 haplotypes (only the European (CEU) panel of 1000 GP was used). Even when the study contains other European, non-Dutch individuals like British and Italians, imputation with GoNL still yields superior results. In another study [68], a negative correlation ($R = -0.81$, $p = 2.5 \times 10^{-4}$) was found between imputation accuracy and genetic distance between the CEU panel of 1000 GP and other European populations. These two findings clearly demonstrate that sequence quality and population affinity in the reference panel can substantially increase imputation results. Finally, improved imputation accuracy was also measured for population-specific reference panels for the Ashkenazi Jewish [40], Sardinian (Italy) and Minnesotan (USA) populations [65], although the last study raised an important concern, which is that due to possible inclusion in the reference panel of samples related to the imputed study, the quality metrics might be inflated for rare variants.

It is important to note that recent comparison studies ⁵ are inconclusive about which tool produces better results [9, 69], for a variety of imputation scenarios. All comparisons have identified marginal cases in which one tool is superior to the other. Yet the differences between time and memory requirements are substantial and are very often the factors determining the final tool choice. Also, as we demonstrated, the computational requirements differ significantly, mainly according to various chunking options [76]. I should also note that existing imputation tools are continuously being improved and new tools being introduced. Developing the most robust, fastest and lightweight tool is in practice an continually active race. As a proof of this, I would urge the reader to compare the titles ⁶ between the papers that introduced IMPUTE2 in 2011 [27] and Beagle 4.1 in 2016 [5]. Therefore, I would advice researchers not to rely on any comparisons that do not include the latest versions of each tool. The latest comparison, to my knowledge (Table 1, [9]), was published in January 2016 and places minimac3 as the best tool in terms of system requirements, whereas all the other tools (minimac2, IMPUTE2, Beagle 4.1) although slower, produce the same quality metrics (mean $R^2$) with an absolute difference of no more than 0.03. Another advantage of minimac3 and Beagle 4.1 is that their source code is open in contrast to IMPUTE2’s closed code.

Finally, apart from the choice of imputation tool, another factor that seems to have a small effect on imputation accuracy is the size of the study panel [83].

---

⁵ for an older comparison of 2010 see Marchini et al. [50], supplementary information S5

⁶ The title of 2011 paper for IMPUTE2 ([27]) is: “Genotype imputation with thousands of genomes”. The title of 2016 paper for Beagle 4.1 ([5]) is: “Genotype imputation with millions of reference samples”.

202
7.1.3 Considerations regarding down-stream analysis

I present three practical considerations regarding the down-stream analysis of imputed studies. These are: (1) estimating the penalty of multiple testing, (2) incorporation of genotype uncertainty in significance tests, and (3) careful alignment of studies before meta-analysis.

Since imputation increases the number of variants in a GWAS the significance threshold should also be adjusted. This threshold is usually adjusted according to the Bonferroni correction method, which states that if $\alpha$ is the significance threshold for a single test, then the threshold that we should apply in cases of $M$ multiple independent tests is $\alpha' = \alpha/M$. Usually $a = 0.05$ for significant association or $a = 0.01$ for highly significant association. Using $M$ as the number of SNPs in a study is too conservative since they are associated through linkage disequilibrium (LD). For this reason, researchers use the number of independent SNPs in a platform as $M$ [71]. For imputed studies, $M$ should be the number of independent SNPs in the reference panel. Current studies raise this threshold to $3 \times 10^{-8}$ for the 1000 Genomes Project [31, 62]. This might make a considerable change in a GWAS design since the significance threshold for an array that contains 0.5 million SNPs (e.g. Affymetrix 5.0) is in the range of $2 \times 10^{-7}$ [43]. Significance thresholds should also be made more stringent for Africans population since the number of independent SNPs is higher due to lower linkage disequilibrium. Therefore, the number of independent SNPs for all samples and for all included population panels in a reference set, should be taken into consideration when applying a significance threshold in an imputed GWA study.

Another relevant consideration is that researchers should always incorporate genotype uncertainty into the significance testing of imputed variants. Namely, if the genotype of an imputed SNP is rounded-up to the one with the highest posterior probability (also called genotype thresholding) then we introduce significant noise into the dataset [61]. For this reason, specially designed tests like SNPTEST’s frequentist methods [51] should be used for association testing of imputed data.

Imputation has also been proven to be valuable in meta-analysis [11]. In imputation meta-analysis, data from multiple studies sometimes from different platforms, can be imputed with the same reference panel. The p-values and effect sizes (usually very small) for each variant in each study can be combined in order to increase detection power and discover novel disease-associated loci. This technique has already been used in order to locate novel, significantly associated variants in complex diseases like Alzheimer’s disease [26], Parkinson’s [57] and in complex traits like cholesterol levels [76]. Li et al. [42] performed an analytic study of the added benefit of imputation on meta-analysis studies. Interestingly, they concluded that possible heterogeneity
between the included studies can introduce uncertainty that will have a detrimental effect on the power of the meta-analysis. These study-heterogeneity factors include different: genotyping platforms, population structures, measurements of phenotypes, SNP filtering (like Hardy–Weinberg equilibrium) and imputation protocols. So careful study alignment is necessary prior to performing a meta-analysis of imputed studies.

7.1.4 Genetic imputation: the future

A final question deals with the future of imputation. Given the sharp decrease in direct sequencing, is Whole Genome Sequencing (WGS) or Whole Exome Sequencing (WES) going to replace the need for imputation? To address this question we need to review the efficiency of the state-of-the-art imputation methods and reference sets.

Visscher et al. [77] compared the number of samples needed to detect a causal variant between a GWAS that contains genotyped samples imputed with a fully sequenced reference panel of the size of HRC and a GWAS that contains WGS samples. The comparison showed that if a causal SNP has a MAF of 0.05% or more, WGS does not offer any significant gain on power of detection. This is impressive considering that they applied conservative estimations regarding imputation quality and also assumed no sequencing errors. The availability of large imputation panels like HRC, raises the bar for the number of samples required to make a WGS GWAS superior to an imputed GWAS; the number rises, 1,000,000 samples if the causal variant has a MAF in the range of 0.001% (1 in 100,000) and a relatively high effect size $\beta$ of 1. If the effect size $\beta$ is lower, then the required sample size (for the same MAF) becomes practically impossible (100 million samples for effect size $\beta = 0.1$). Therefore, I conclude that imputation will continue to be a major part of GWAS studies, assuming that the reference panels will continue to grow and to include additional populations.

This conclusion is also backed up by a cost-effect analysis. Assuming a WGS price per sample at $1000 and a genotyping cost of $50, then with the same budget an imputed GWA study is 13 times more powerful to detect a common variant and 4 times more powerful to detect a rare variant than a WGS GWAS for a single-variant-based association analysis [82]. Interestingly, imputation can even help to increase the cost-effectiveness of sequencing-based GWAS. As it has been demonstrated [63], imputation of very low-coverage, and therefore inexpensive, sequencing data (average coverage: 0.1-0.5) can increase the power of a GWAS. The same effect has been observed when SNP array data are combined with exome chip data in the study panel, prior to imputation [36]. As a consequence, when budget is constrained, it is preferable to sequence more individuals at lower coverage than a lower number at higher coverage [46]. Surprisingly, the same stands for SNP array studies. Namely, a GWAS that contains
imputed samples from a low density SNP array is more powerful than a GWAS as the same cost performed on imputed samples from a higher density array (comparison between HumanHap 300 and HumanHap 1M arrays[2]).

Of course, methods based solely on WGS and WES data are not always deemed inferior for the identification of novel causal variants compared to imputation. Imputation has become an essential accompaniment to the general GWAS family of methods. The main objective of GWA studies is to locate associated variants for a disease that will provide insights into the causal variant through linkage. Whole genome sequencing and exome sequencing techniques, on the other hand, have a predominantly different architecture since they do not rely on linkage for identifying the causal variant but assume that a direct (and qualitative) assessment of its sequence will be acquired[74]. The main consideration in this approach is the large amount of human genetic variation [16] and the plethora of candidate variants. To tackle this, we assume that complex diseases are guided by a spectrum of very low frequency variants with deleterious effects [35, 56]. These variants are segregated in the population and their aggregate effect remains undetected in GWA studies for two reasons. The first is that, in most GWAS, each variant is tested independently, and the second reason is that very rare variants are filtered out as part of a quality check step. Therefore, both imputation-based GWAS and sequencing techniques have a bright future in the race to revealing causal variants for complex diseases.

Another benefit of imputation is that it saves us from a lot of data storage and management. This is a major benefit since, according to estimations, by 2025 the need for storage of genetic data will surpass that of popular websites like YouTube and Twitter [73] and will reach the astronomical size (or better.. genomic size?) of 2-40 Exabytes (1 Exabyte is $10^{18}$ bytes). Despite the decrease in sequencing costs, the technical issues that arise from efficiently managing and analyzing genomic data on such a scale will make data-efficient methods like imputation much more favorable in the research community. Moreover, direct genotyping techniques are faster, generate fewer and more qualitative data, and are already widely commercialized. Additionally, the existence of open, good quality repositories of clinically important variants and the tools7 to locate them in commercial DTC (Direct To Customer) genotype data points to the bright future for this area. Besides, imputation is already being used to increase the screened variants from DTC genotype data8.

Perhaps a point of concern is the current low availability of imputation reference

---

7http://promethease.com/

8See blog post from Peter Cheng and Eliana Hechter: “Learning more from your 23andMe results with Imputation”: http://www.genomesunzipped.org/2013/03/learning-more-from-your-23andme-results-with-imputation.php
panels for large parts of the human population, mainly from developing countries. Despite the admirable effort of the 1000 Genomes Project Consortium to include populations from a wide and diverse set of human populations, this effort has not created the expected momentum. Today the largest reference panel for African studies contains 883 samples [53] which is rather low given the large genetic heterogeneity of this population. Most Asian studies are still using panels from the 1000 Genomes Project [18] which contains fewer than 1000 samples for these populations. In contrast, modern reference panels for studies with European ancestry contain in the order of 30,000 samples. This phenomenon has also been observed in GWA studies [66]. Whether or not this uneven distribution is temporary, is a question to be answered in the future. An optimistic view is that novel reference panels are continuously being constructed for national [34] and regional [78] populations, or even for certain genomic regions (e.g. HLA [59]). So we might reach an equilibrium in the future, when all human heterogeneity can be explored in high resolution through imputation. Nevertheless, given the trends of increasing human travel and migration [10] we should expect to see a consequent rise in admixed populations. Therefore, instead of many population-specific reference panels, I advocate the use of unified reference panels that include: (1) as many different populations, (2) as higher number of haplotypes and (3) as many variants as possible [5].

7.2 Integration as a vehicle towards clinical genetics

The majority of the work in this thesis has concentrated on a central problem; how to offer simple, yet powerful solutions that integrate a variety of tools that are essential for analysis in bioinformatics. The visionary goal of this effort is the realization of the scope of “translational bioinformatics” [1]. Translational in the sense that all findings and discoveries from a diverse area that spans molecular biology and genetics can be interpreted in clinical knowledge and applications, and therefore used for diagnostic and therapeutic purposes. This is a crucial component of the new and widely envisioned field of “personalized medicine” [60].

Of course, just before reaching clinical applications of new genetic knowledge lie two fundamental milestones. The first is understanding of inter-individual genetic heterogeneity and the second is unraveling the complexity of genome-phenome interactions. So where do we stand today on the road towards these milestones? For sure, we are still far off having causal or mechanistic models at our disposal that describe how genetic regulation affects the phenotype (although there is lively discussion on whether such models actually exist [49]).
Nevertheless, we do have high quality data that describe nearly all stages of genetic regulation and transcription for various cell types, diseases and populations. DNA sequences, RNA sequences and expression, protein sequences and structure, small molecule metabolite structures are all being routinely measured in unprecedented numbers nowadays\cite{15}. Moreover, prospects for data collection seem very bright as the throughput rates are increasing, quality metrics are improving, and prices are dropping.

The maturity, quality and efficiency of existing bioinformatics tools has risen tremendously in the last few years. This can be attributed to the increasing emphasis on IT literacy and skills in biology and genetics curricula \cite{3}, the greater inclusion of professional IT developers in biology labs, and, of course, on the intrinsic computational challenges of modern genetic analysis.

This brings us to the conclusion that the integration is the intermediate step that is missing between the current situation, which is the availability of enormous data collections coupled with qualitative tools, and the objective, which is mechanistic or causal models.

**7.2.1 The Roadmap**

It is evident that pipelines per se cannot provide sound solutions to the problem of modern genetic analysis despite their sophistication and technical completeness. The driving force of competent pipelines lies with the developers and with the maintainers of their components. Therefore, the philosophy that guides the development of these components should be characterized by altruism, extrovert thinking, and having novice users as the main target group. Although these characteristics seem abstract and overambitious, we have presented concise and easily applicable development guidelines in Chapter 2, that can help make them a tangible reality.

From a more technical perspective, integrating existing tools and data in bioinformatics is basically the process of resolving a series of considerations. In Figure 7.1, I present a visualization of these considerations. At the core of modern bioinformatics lie four basic components: tools, data, scripts and workflows. Each of these components can (and should) be extended in a four-dimensioned space. These dimensions are Documentation, Wrapping, Composition and Collaboration.
Figure 7.1: Integration can be visualized as a four-dimensional space in which we can extend existing tools, data, scripts and workflows (called components). In each of these dimensions there are specific issues that a developer or maintainer should consider. Some examples are: (1) **Documentation**, do we describe all the components sufficiently? Have we included tests and examples? (2) **Wrapping**, do we wrap all dependent components in a simple object that we can easily deploy in a HPC? (3) **Composition**, is it easy to connect our tool with other components for up-down and side-ways analysis? and (4) **Collaboration**, is it easy for other users to find, edit, comment and rate our component? Of course the set of considerations that lie in each dimension (ovals) is rather indicative and far from complete. What is more important is that, since this is a multi-dimensional space, each consideration in any dimension also affects all the other dimensions. For example: does our virtualization method have adequate documentation (Wrapping + Documentation)? Have we allowed other users to extend our semantic description (Composition + Collaboration)? etc.
Documentation considerations refer to how to extend existing components by providing simple and easy-to-follow directions that cover all the details of complex bioinformatics tasks. These details should cover system requirements, measurement of input/output quality, common pitfalls and ways to troubleshoot them. This was demonstrated in Chapter 3 with a break-down and detailed analysis of a modern, genotype imputation pipeline.

A second consideration is Wrapping, which is how to provide ‘one-stop’ solutions that wrap complex pipelines, hide unnecessary implementation details and target researchers that ‘just want to do this task now’. As I demonstrated in Chapter 4 with ‘MOLGENIS-Impute’, it is possible to provide open source, documented, highly customized and directly executable pipelines that conduct the complete analysis. These wrappers are HPC-aware and tweak the pipeline according to the specifics of the execution environment at the researcher’s disposal. These wrappers are also excellent components in auxiliary pipelines since they can be easily inserted into an arbitrary analysis.

Collaboration considerations refer to how to complement existing components by exploiting the ‘power of the masses’. This is another crucial source of innovation that can be harvested by simply being open to and accepting of feedback from an active community. Open, receptive and crowdsourced environments like PyPedia described in Chapter 5, can bring together specialists with a diverse skillset and interests. These domain-agnostic environments can create qualitative solutions that gradually emerge in unprecedented ways out of user collaboration.

The fourth consideration is Composition. Pipelines should be aware of the limitations of the included components and should also help to tackle problematic implementations or incomplete user input. As demonstrated in Chapter 6 with MutationInfo, pipelines should build synergies between incomplete tools rather than simple sequential analysis pipelines.

7.2.2 The benefits

One last question is what is the immediate societal benefit from making these characteristics (altruism, extrovert thinking, and targeting novice users) a reality? The answer lies in making clinical genetics an essential part of everyday medical practice in modern healthcare services. The first step has already been taken in the form of the official recognition of the clinical/medical genetics specialty as an EU-wide medical specialty in 2011. Sixteen European countries (including the Netherlands) already have national

\(^9\text{Clinical/medical genetics officially recognised as an EU-wide specialty! }\text{https://www.eshg.org/111.0.html}\)
education schemes that fit the official description of the curriculum that is necessary to obtain recognition in this specialty [47]. Medical and clinical genetics are expected to improve diagnosis, help health providers to take more informed decisions, develop drugs tailored to an individual’s genetic profile, and inform high risk groups of their possible disease predispositions. News articles and case study publications are filled with “success stories” that belong to one or more of the above categories. Although the maturity, validation and experimental status of these efforts vary, it indicates that we are getting closer to the systematic adoption of clinical genetic practices on a wide scale. Of course, in the more distant future, technologies like CRISPR/Cas gene editing, cancer immunotherapy [8], and the use of pocket-sized sequencers [24], which are today still in their infancy, will revolutionize healthcare even further.

Thus, we can finally be confident that clinical genetics is today being routinely applied through tests that deliver the correct diagnosis for the majority of Mendelian (or single-gene) diseases [67]. Most of these tests are applied as a follow-up step, after a positive result from a newborn screening program. Yet the picture is more oblique in cases of complex disorders, in which a large set of genes or other regulatory elements are implicated. Although studies have identified hundreds of thousands of associated variants [39], the estimation of the pathogenicity of each variant and the elucidation of their mechanisms of actions are still an area of active investigation for most complex diseases. However, this information is essential in a clinical setting in order to make a correct risk assessment and subsequently to apply optimal therapeutic strategies [33, 64, 49]. Francis S. Collins, when he attempted to assess the long-term effects of the Human Genome Project in 2001, among others, predicted that “By 2020, it is likely that every tumor will have a precise molecular fingerprint determined, cataloging the genes that have gone awry, and therapy will be individually targeted to that fingerprint” 10. Of course, research projects like the Cancer Genome Atlas (TCGA) [80] have set out to realize this prediction but we now see that timeframe was certainly overestimated. This is not because prominent researchers like Francis S. Collins were over-ambitious at that time but it is mainly due to the fact that the more light we shed on genetic regulation, through improved sequencing or other mass profiling techniques, the more complex it is revealed to be. For example, today there are more than 20 kinds of -omics data, most of them unknown to geneticists 20 years ago 11. Yet, today we can be

10It is also worthwhile to review today the major challenges of genetic research that was compiled by the same author in 2003 [7]. The challenge that refer to clinical genetics (titled Grand Challenge II-5) could very well repeated in a today’s research article. Another study that attempts to predict the future at 1999 is [41]

confident that we have finally reached the point where the wealth of available data [52], the efficiency of existing technologies, and the collaborative mindset of modern scientific society [22] is adequate to produce medically applicable knowledge even for complex diseases [25]. This is possible, although only by following a very specific course of action, i.e. integration [21, 55, 79]. Therefore, as shown throughout this thesis, I can confidently state that by integrating the vast amounts of data now available, modern analysis methods and the current standards of genetics research, we can further enrich medicine and its practice, and consequently improve public health.
Bibliography


