Marina Mustonen

Software tools in medical genetics: a systematic mapping study

Master's thesis of mathematical information technology

December 18, 2019

University of Jyväskylä

Faculty of Information Technology

Author: Marina Mustonen

Contact information: marina.s.mustonen@gmail.com

Supervisors: Ville Isomöttönen

Title: Software tools in medical genetics: a systematic mapping study

Työn nimi: Lääketieteellisen genetiikan digitaaliset työkalut: systemaattinen kirjallisuuskartoitus

Project: Master's thesis

Study line: Software Engineering

Page count: 40 + 14

Abstract: Rise of commercial gene tests and whole-genome sequencing becoming easier and cheaper brings many benefits to health care, but also challenges. The challenges include storing, sharing and managing huge amounts of data, analyzing genetic data to find relevant information, and interpreting and visualizing complex genetic data. The issue of data privacy is also a major concern in medical genetics. Efficient and well-designed software can help with these challenges. In my thesis, I strive to provide a systematic mapping of the literature on medical genetics software and answer questions about what types of research approaches and technological focuses there are in scientific articles on the subject, is data privacy addressed in the articles, and what types of journals the articles are published in. The results show that validation is the most common research approach. This result is possibly partly due to the inclusion criteria and categorization approach used. Preprocessing and analysis of genetic data is the most common technological focus in the papers, likely due to genetic data and analysis of it being very complex and requiring different kinds of software to help with it. Data privacy is not addressed in most of the papers, which could partly be due to it not deemed an issue in all types of software. Most of the articles were published in journals related to medical genetics, which is probably due to the genetics professionals using the software reading more journals related to their field.

Keywords: Medical, Genetics, Software, Genome, Health care, Gene test, Big data.

Suomenkielinen tiivistelmä: Kaupallisten geenitestien tarjonnan kasvu ja kokonaisten genomien sekvensoinnin muuttuminen yhä helpommaksi ja halvemmaksi mahdollistavat geneettisen tiedon hyödyntämisen lääketieteessä. Tästä on paljon hyötyä terveydenhoidossa, mutta siihen liittyy myös haasteita. Haasteita tuovat mm. geeniteknologian tuottamien suurten datamäärien varastointi, jakaminen ja käsittely, geneettisen datan analysointi niin, että oleellinen tieto saadaan selville ja monimutkaisen geneettisen datan tulkitseminen ja visualisointi ymmärrettävässä muodossa. Geneettisen datan suojaaminen on myös yksi suuri haaste lääketieteellisessä genetiikassa. Tehokkaat ja hyvin suunnitellut digitaaliset työkalut voivat auttaa näissä haasteissa, ja uusille työkaluille onkin kasvava tarve, kun lääketieteellinen genetiikka yleistyy. Tässä tutkimuksessa pyrin tekemään systemaattisen kartoituksen lääketieteellisen genetiikan digitaalisista työkaluista ja vastaamaan kysymyksiin siitä, millaisia tutkimuksia aiheesta on, millainen teknologinen fokus niissä on, onko datan suojaamista käsitelty artikkeleissa ja minkä tyyppisissä tieteellisissä lehdissä artikkeleita on julkaistu. Tuloksista selviää, että suurin osa artikkeleista on validointitutkimuksia. Tämä tulos johtuu luultavasti osittain siitä, miten kriteerit artikkelin mukaan ottamiseen on määritetty ja miten kategorisointi on tehty. Suurin osa artikkeleista käsittelee geneettisen datan esikäsittelyä ja analysointia, mikä johtuu luultavasti siitä, että geneettistä dataa on niin monenlaista ja sen analysointi on vaikeaa, jolloin on tarvetta monille erilaisille digitaalisille työkaluille auttamaan siinä. Datan suojaamista ei käsitellä suurimmassa osassa artikkeleita, mutta se johtunee ainakin osittain siitä, että datan suojaamisen ei katsota olevan oleellista kaiken tyyppisissä digitaalisissa työkaluissa. Suurin osa artikkeleista oli julkaistu lehdissä, jotka liittyvät lääketieteelliseen genetiikkaan (genetiikka, lääketiede ja bioinformatiikka) ja julkaisuja ei ollut juurikaan yleisemmin tietotekniikkaa käsittelevissä lehdissä. Tämä johtunee siitä että, lääketieteellisen genetiikan digitaalisia työkaluja käyttävät luultavasti lähinnä alan ammattilaiset, jotka lukevat enemmän genetiikkaan liittyviä julkaisuja kuin yleisesti ohjelmistoihin keskittyviä lehtiä.

Avainsanat: Lääketiede, genetiikka, ohjelmisto, sovellus, digitaalinen työkalu, genomi, terveydenhuolto, geenitesti, sekvensointi

List of Figures

List of Tables

Contents

1 Introduction

It is becoming more common to use genetic data in health care, to make more precise diagnoses, to asses risk of disease more efficiently and to prescribe drugs that are more suitable for each individual (Shirts et al. 2015, Evans et al. 2016, McGrath and Ghersi 2016). For example, treating cancer has benefitted greatly from using genetic data (Chang 2018). Health care professionals have often not had much training in using genetic data, so analysing and interpreting the data is often left to specialists in genetic field (McGrath and Ghersi 2016). Efficient software can help make genetic data more available for health care professionals and help lighten the workload of specialist, as well as aid with other challenges facing the use of genetic data in health care (Zhang et al. 2018). There is need for systematic mapping of the literature on this subject, which this thesis aspires to give.

1.1 Brief introduction to human genetics

Description of human genetics can be found e.g. from a book by Pasternak (2005). Hereditary material, which allows traits to be passed from parents to their children, is contained in deoxyribonucleic acid molecules (DNA). The organization of DNA in the cell is described in Chapter 3 of the book (Pasternak 2005). DNA is coiled around histone proteins to form chromatids and two pairs of them are joined with a centromere to form a chromosome. Chromosomes are located in the nucleus of the cell. Humans have 22 pairs of chromosomes called autosomes and two sex-determining chromosomes (females have two X chromosomes and males have one X and one Y chromosome). One chromosome of each pair is inherited from the mother and one from the father. Chapter 3 of the book (Pasternak 2005), as well as Read (2017) give definitions for the terms genotype and phenotype. The genetic identity of an individual (i.e. their genetic structure) is called a genotype (the term is sometimes used also to describe particular gene or set of genes). Phenotype is the observable characteristics or traits of an individual, and it is affected by genotype, as well as environmental factors and epigenetics (heritable changes in expression of genes not involving changes in DNA sequence).

Chapter 4 of the book (Pasternak 2005) describes the structure of DNA. DNA is composed of units, which have a sugar-phosphate and a base. The base is either adenine (A), thymine (T), guanine (G) or cytosine (C). A and T or G and C form pairs with weak chemical bond, which is why DNA is double-stranded and forms a double helix shape. Figure 1 (Modified from Wikimedia Commons, accessed Nov. 19, 2019, original from www.genome.gov) shows the structure of DNA and its organization on different levels.

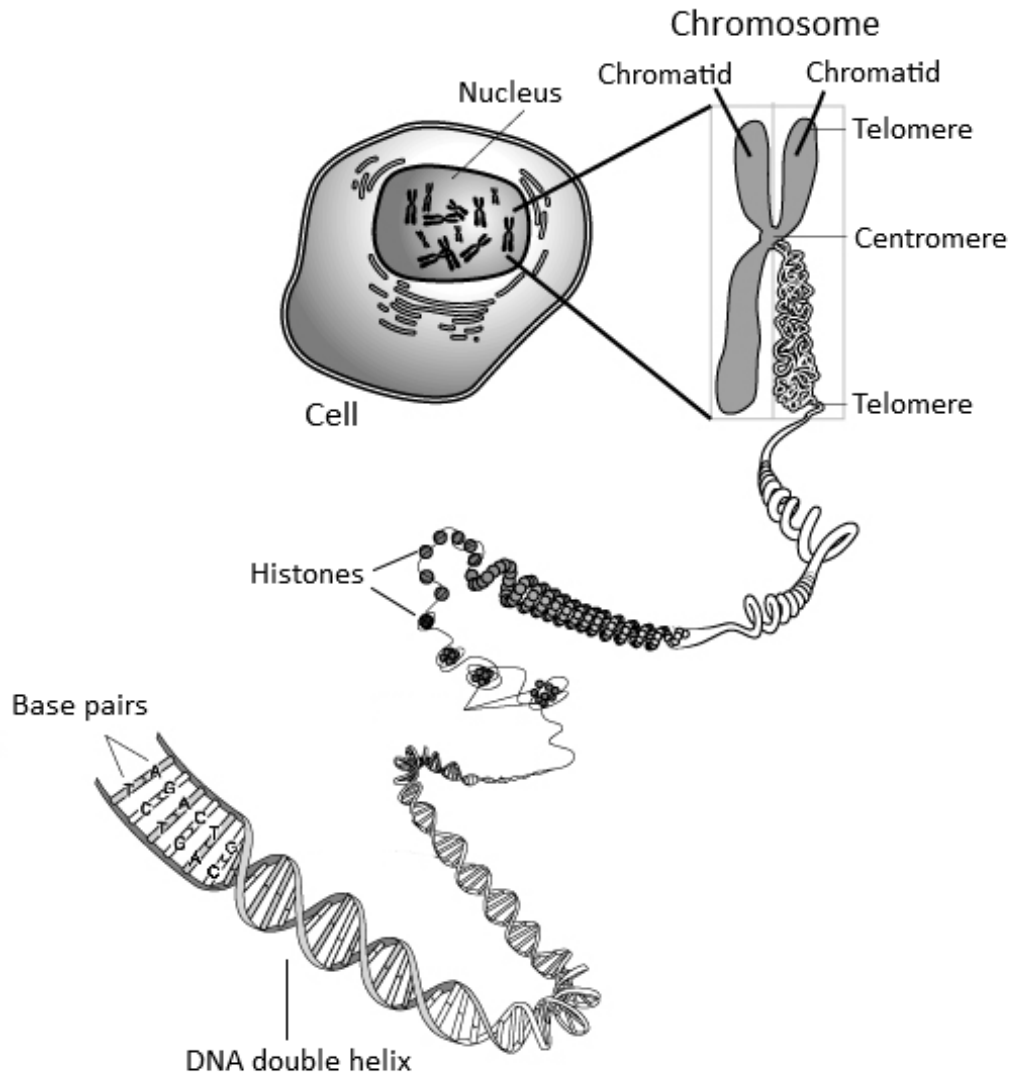

Figure 1. Structure of DNA and organization into chromosome (Modified from Wikimedia Commons, accessed Nov. 19, 2019, original from www.genome.gov).

Chapter 4 of the book (Pasternak 2005), as well as Read (2017) also describe genes and their transcription and translation processes. Genes are segments of DNA that hold the information for (usually) one protein each. Genes have parts that code for the protein and non-coding parts that can have other functions, e.g. in regulating gene expression. When the protein that the gene codes, is needed, cell gets a signal to start transcription. It is a process, in which the DNA strand of the gene is copied into messenger RNA (mRNA). RNA is ribonucleic acid, which is similar but not identical to DNA. The double helix unwinds on the location of the gene on the chromosome and one of the strands is used as a template to form a strand of RNA. The formed strand is spliced so that only the protein-coding parts remain, resulting in mRNA, which is used to synthesize the protein. Alternative splicing of the original RNA can result in different proteins, which means that long-held believe of "one gene-one protein" is not accurate. Protein is synthesized in a process called translation. Each three-base segment of the mRNA is called a codon and it corresponds to a specific amino acid, which are the units of proteins. The mRNA is used as a template to add the amino acids one by one to form the protein. The phases in transcription and translation are shown in Figure 2 (Wikimedia Commons, accessed Nov. 19, 2019).

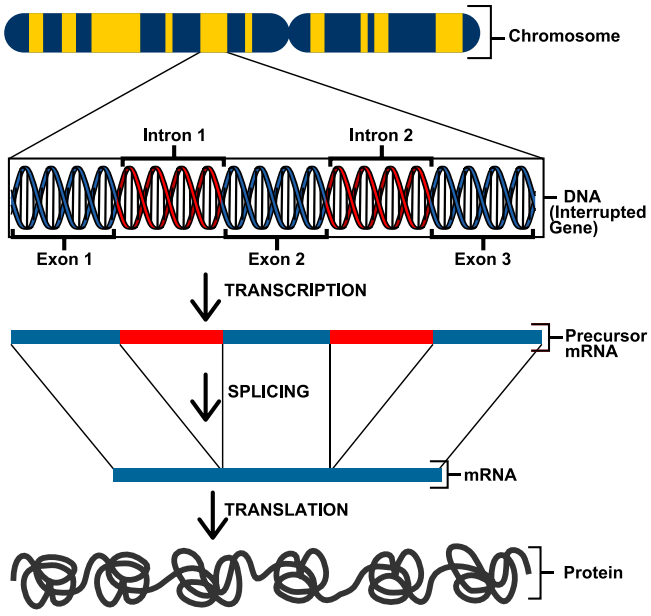

Figure 2. DNA transcription to mRNA and mRNA translation to protein (Wikimedia Commons, accessed Nov. 19, 2019).

Process of replicating DNA and the possible mistakes in it are described in Chapter 4 of the book (Pasternak 2005), as well as by Besenbacher and colleagues (2016) and by Read (2017). Before a cell divides, DNA is replicated. Strands of DNA are separated and used as a template for the new strands. Each new base is paired with the template base. Sometimes there are mistakes in pairing the bases, which are called mutations. Replacement of just one base with another is called a point mutation and it can change the protein that is coded or affect the regulation of the gene. However, more often point mutations do not affect the genes function since some changes do not affect the amino acid that is coded, or changing some amino acids do not affect the function of the protein. Other types of mutations are insertions (piece of DNA is added), deletions (piece of DNA is removed), duplication (piece of DNA is copied twice), frameshift mutations (proteins are coded in segments of three bases, thus addition or deletion of one base changes the whole reading frame) and repeat expansion (there are short repeats in DNA sequences and changes in the number of repeats can affect the protein that is coded).

1.2 Effect of genetics on health

Effects of mutations on health are described on Chapter 4 of the book by Pasternak (2005), as well as by Besenbacher and colleagues (2016) and by Read (2017). Mutations are basis of all variation in genotypes and enables evolutionary adaptation, but many mutations that affect the gene function are also connected to various disorders and diseases. Mutations are passed on to new strands when using the strand with the mutation as a template. Mutations happening in somatic cells are not passed on to offspring, so they only affect the individual in which the mutation happened in, but mutations in the germ line cells (cells that sperm and egg cells are formed from) are. Most mutations are recessive, meaning that the effect is masked if the corresponding gene in the complementary chromosome is functional. Effect of recessive mutation can become apparent when a child inherits a defective gene from both parents that are not themselves sick because they have a functional copy also.

In addition to mutations, also changes in chromosomal levels are of interest in medical genetics. Chromosomal level changes are described in Chapter 2 of the book (Pasternak 2005), as well as by Khandekar and colleagues (2013) and by Read (2017).There can be changes in

the chromosome number (aneuploidy), in which case the individual can have an extra chromosome or miss one. There can also be structural changes in the chromosome, like deletions of parts of the chromosome, insertions of extra pieces or inversions of part of the chromosome (segment of the chromosome has flipped 180 °). Each chromosome has many genes in it, so most chromosome level changes, especially aneuploidy and large-scale deletions, are lethal, with some exceptions.

All changes to the DNA, e.g. copy number variation of the gene, repeat variation, point-mutations, deletion and insertions, as well as chromosomal changes, can affect how the gene functions. Gene might not be expressed (i.e. produce the protein it is coding) properly, or mutations in the regulatory parts may lead to overexpression of the gene, which can also have harmful effects (Pasternak 2005 Chapter 2, Khandekar et al. 2013, Read 2017). In addition to structural changes in the DNA molecule, methylation, which is a process in the cell that can make some parts of the DNA unavailable for use thus affecting the expression of the genes, is of interest in medical genetics (Costello and Plass 2001, Singh et al. 2003). Looking at the gene expression instead of the structural changes to the DNA, can be used as a diagnostic tools, although finding the root cause of the expression differences, e.g. mutation or methylation of the DNA, can affect how widely usable the information is (Hedenfalk et al. 2001, Stranger et al. 2005).

1.3 Application of genetics in health care

Using genetic information in health care can lead to more efficient and precise diagnoses and treatment (Abel et al. 2005, Shirts et al. 2015, Evans et al. 2016, McGrath ja Ghersi 2016, Miller et al. 2017). Genetic data is often used to find connections between genotype, which is persons genetic makeup, and phenotype, which is persons observable characteristics, to find connections between genomes and diseases (direct or increase in the risk of a disease), or between genomes and drug-resistance (Brookes and Robinson 2015). Finding those connections helps with diagnosis, risk assessment and treatment of a disease, e.g. choosing the right kinds of drugs. Genetics can be used in health care by looking at variations (e.g. single nucleotide variation, repeat variation or copy number variation) in genes or regulatory elements, methylation of genes or gene expression data (Hedenfalk et al. 2001,

Stranger et al. 2005, Klonowska et al. 2015). Targeted genetics tests, which look at specific, known disease-linked genes, can help with diagnoses, risk analyses and treatment (McGrath and Ghersi 2016), but with whole-genome and whole-exome (exome includes only the protein coding parts) sequencing becoming easier and cheaper to do, it is becoming more common to look at the whole genome of a patient and try to find links between diseases and genes (Abel et al. 2005, Rabbani et al. 2014, Brookes and Robinson 2015). Many gene-disease associations are already known, but a lot is yet to be discovered also.

Targeted disease prevention and treatment approach is often called precision medicine. It includes taking genetic, lifestyle related and environmental factors into consideration when preventing, diagnosing or treating diseases (McGrath and Ghersi 2016). Lifestyle and environmental factors have been taken into consideration for a long time in medicine, whereas genetics in medicine is relatively new and presents many challenges (McGrath and Ghersi 2016). One of those challenges is how to store, manage and share the huge amount of data generated with genetics techniques like specialized gene tests and more comprehensive mappings of the genome, e.g. whole-genome sequencing (Shirts et al. 2015, Reali et al. 2018). Another challenge is how to analyze the data to find relevant information, e.g. about essential disease-gene or drug-gene relationships (Milicchio et al. 2016). Yet another challenge is how medical professionals, who are often lacking in genetics training, can interpret and present to the patients the results of e.g. gene tests or sequencing data (Tinkle and Cheek 2002). Efficient software tools can be used to make these challenges easier (Gobalan and John 2016, McGrath and Ghersi 2016, Milicchio et al. 2016, Reali et al. 2018, Zhang et al. 2018).

There is also the challenge of keeping genetic data private, both in medical genetics generally and in developing software to help with using genetic and sharing genetic data in health care (Fuller et al. 1999, Abel et al. 2005, Reali et al. 2018, Thorogood et al. 2018). Developing new software to help utilize genetic data in health care brings many benefits in combating diseases, but genetic data is highly personal data and care need to be taken to ensure privacy of the patients and anonymity of the genetic data. Genetic data can be misused for insure and employment discrimination or e.g. in custody cases (Fuller et al. 1999, Abel et al. 2005). Especially storing and sharing data in databases as well as web-based platforms for analysis

and interpretation of data could present security risks if not designed carefully (Reali et al. 2018, Thorogood et al. 2018).

1.4 Software tools in medical genetics

Acquiring genetic data is relatively easy but analyzing and applying it to medical care is more complex. Many different databases have been established for the vast amounts of genetic data generated, ranging from general databases (e.g. GenBank) to databases dedicated to specific disease or group of diseases (e.g. cancer databases or databases for rare diseases) and national databases (Brookes and Robinson 2015). Databases help store and share the genetic data, but to make the data easier to apply there is also need for software tools for structuring and mining the data (Thorogood et al. 2018). After sequencing a gene, it needs to be annotated, i.e. genes location identified and genes function determined. To do that, and for people to be able to access that data easily and compare their own genetic data to it, many software tools have been developed, e.g. VarWatch (Fredrich et al. 2019) and Pharm CAT (Sangkuhl et al. 2019). To find associations between gene variants and diseases or drugs, the variants need to be identified and large amounts of genetic data from different individuals need to be analyzed. Software tools for analysis and e.g. web-based tools for integrating those tools to databases have been developed (Zhang et al. 2018). Machine learning technologies can be used to predict variation and the effect of the variation (Reali et al. 2018, Zhang et al. 2018). Examples of machine learning software for medical genetics include FATHMM-MKL, which predicts the function of both coding and non-coding regions of DNA (Shihab et al. 2015) and hyperSMURF for predicting non-coding variants associated with rare diseases (Schubach et al. 2017).

Even with the help of software, gene annotations, variant determination, analysis and finding gene-disease associations are usually done by professionals who are trained to handle genetic data (McGrath and Ghersi 2016). Interpreting the results, even for basic gene test, but especially for whole-genome sequencing, also often requires specialists trained in the field. However, interpretation and visualization of genetic data can be made easier for regular health care professional with software tools (Brookes and Robinson 2015, Shirts et al. 2015, Evans

et al. 2016, McGrath and Ghersi 2016). For example, Klonowska and colleagues (2015) review several portals used for visualizing and interpreting cancer gene data.

1.5 Systematic mapping study

Systematic literature mapping is a way to provide a general overview of a specific research area. Petersen and colleagues give guidelines for performing a systematic mapping study in the field of software engineering in their 2008 paper and expand and elaborate them in their 2015 paper. According to those guidelines, in systematic mapping study a protocol with several steps is followed to categorize published literature on the research area and to provide a summary, often in visual form, of the results. A general overview of the mapping process is presented in Figure 1 (Petersen et al. 2008).

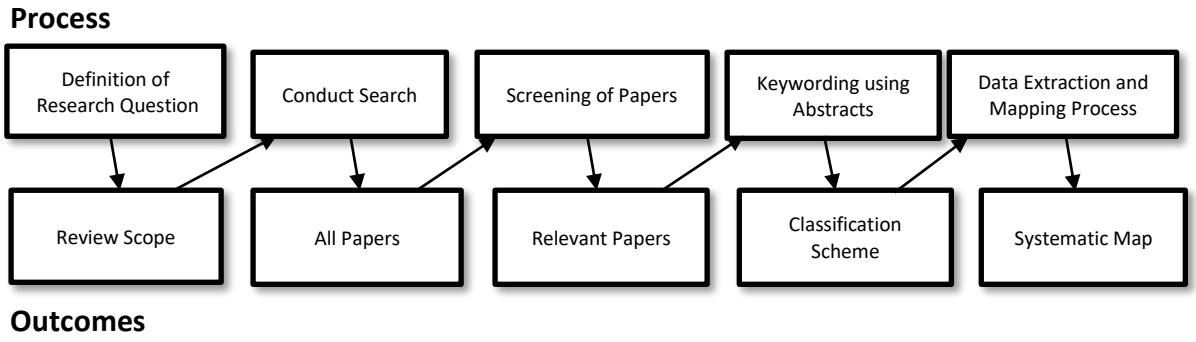

Figure 3. The systematic mapping process (Petersen et al. 2008).

According to Petersen and colleagues (2008 and 2015), the process starts with defining the research questions and the scope of the research. In a mapping study, the main research question is often quite broad and about what is known about a specific topic. Higher level question can be broken down to several more specific questions to help with the data extraction. Since a general overview is the aim, the research questions are not supposed to be highly specific.

Following the guidelines set by Petersen and colleagues (2008 and 2015), after the research questions and scope are defined, a search string to use on scientific databases or a process

for going through relevant sources manually is defined. According to Petersen and colleagues (2008 and 2015), the research questions drive how the search is done. The search for relevant papers is conducted, which results in papers to be screened on the next step.

According to Petersen and colleagues (2008 and 2015), the papers are screened by using inclusion and exclusion criteria. Inclusion and exclusion criteria are defined based on what can help answer the research questions. Relevance of the topic is one criteria, but there can also be criteria based on e.g. certain time period, language of the papers or publication venue. The aim is that after the screening only relevant papers remain.

Following the guidelines set by Petersen and colleagues (2008 and 2015), after finding the relevant papers, data extraction and classification is done to them. Petersen and colleagues (2008 and 2015) describe that one way to do it is keywording the abstracts i.e. looking for keywords and concepts in abstract and using them to find representative categories. Categories can be based e.g. on the topic or type of contribution, but a more general way is to categorize based on research approach. Petersen and colleagues (2008) summarize a way to categorize the papers based on the research approach used in the papers. The categorization is based on paper by Weiringa and colleagues (2005). The categories are presented in Table 1. In their 2015 paper, Petersen and colleagues expand on the categorization criteria, especially on the difference between validation and evaluation research. They specify that validation is not used in practice whereas evaluation is done in a real-world context.

Category	Description
Validation Research	Techniques investigated are novel and have not yet been implemented in practice. Techniques used are for example experiments, i.e., work done in the lab.
Evaluation Research	Techniques are implemented in practice and an evaluation of the technique is conducted. That means, it is shown how the technique is implemented in practice (solution implementation) and what are the consequences of the implementation in terms of benefits and drawbacks (implementation evaluation). This also includes to identify problems in industry.
Solution Proposal	A solution for a problem is proposed, the solution can be either novel or a significant extension of an existing technique. The potential benefits and the applicability of the solution is shown by a small example or a good line of argumentation
Philosophical Papers	These papers sketch a new way of looking at existing things by structuring the field inform of a taxonomy or conceptual framework
Opinion Papers	These papers express the personal opinion of somebody whether a certain technique is good or bad, or how things should been done. They do not rely on related work and research methodologies
Experience Papers	Experience papers explain on what and how something has been done in practice. It has to be the personal experience of the author.

Table 1. Research Type Facets (Petersen et al. 2008).

Following the guidelines set by Petersen and colleagues (2008 and 2015), the final step is mapping the results to a systematic map. Numbers of articles in each category are counted and presented. Often the results are presented in a visual form, e.g. with bubble plots or bar plots.

2 Materials and methods

2.1 Background information

To my knowledge, a systematic mapping study has not been done for medical genetics software. I searched Google Scholar, PubMed and Scopus with "systematic mapping study" AND "medical genetics" AND software. There were three results on Google Scholar. One of the results was a Master's thesis and not in English, one was clearly not about a systematic mapping study and one was a systematic mapping study, but not specifically about medical genetics software, but about feature-selection techniques used in big genomic data analysis. In PubMed there were 18 results but none of them were actually about systematic mapping study of medical genetics software. In Scopus there were three results also and none of them were a systematic mapping study on medical genetics software.

Medical genetics is a fast-growing field with demand for more precise and effective software increasing. There is need to map what the situation is regarding the literature covering the current software in medical genetics.

2.2 Research questions

In this thesis, I strive to give a general view of the literature about medical genetics software. First, I made a more general (i.e. not tied to the topic) categorization based on the research approach used in the paper, as summarized by Petersen and colleagues (2008). I followed Table 1 categorization and use Petersen and colleagues later paper (2015) to help with deciding which categories each paper goes to. Then I categorized the literature I found according to the technological focus i.e. the type of aid they provide in the field of medical genetics. Technological focus categories were determined based on the initial reading on the topic and looking at what types of technological focuses arose from that. One category is data storage and sharing, e.g. databases, electronic records and other storage solutions (henceforth referenced as Storage). Second category is preprocessing and analysis of the data (henceforth referenced as Analysis). The third category is visualization and interpretation of the data

(henceforth referenced as Interpretation). I also looked into if the issue of data privacy is addressed in the included literature. In addition, I took a look at what types of journals the papers are published in. Journal types were determined by looking what kinds of journals the articles were published in and making logical groups based on the focus of the journals. Thus, the research questions are:

1. What types of research approaches there are in the papers?
2. What is the technological focus (Storage, Analysis or Interpretation) of the papers?
3. Is data privacy addressed in the papers?
4. What types of journals the papers are published in?

2.3 Sources of literature

Literary search was made on several sources of literature. The sources are Google Scholar, IEEE Xplore, PubMed, ISI Web of Knowledge, Scopus and ACM Digital Library. These sources provide good coverage of the subject, Google Scholar, ISI Web of Knowledge and Scopus being more general databases of literature, whereas IEEE Xplore and ACM Digital Library are more technical oriented databases, giving good coverage of the software-part of the mapping. PubMed is biomedical database, giving coverage to the medical genetics –part of the mapping.

2.4 Refining the search terms

Test searches were done on some of the literature databases to refine the search terms. Searches for “since 2015” (Google Scholar and ACM Digital Library) or “2015-2019” (IEEE Xplore and Scopus) or the last 5 years (PubMed and ISI Web of Knowledge) were used. Also, in IEEE Xplore results were restricted to the “Journal” type (since it was an option to exclude the ones that were not needed, i.e. Conference, Magazines and Courses), in ISI Web of Knowledge only Articles and Reviews were included (Proceedings and Editorial Notes were excluded) and in ACM Digital Library only Periodical were included (e.g. Proceedings were excluded) to get only scientific articles in journals. Table 2 shows the date and search words used in the test searches, and in bold the actual search that was performed.

Google Scholar			
Date	Search terms	Result	Comment
17.9.2019	Medical AND genetics AND software	233 000	Too general, need to refine more
17.9.2019	Medical AND genetics AND software AND gene testing	15 500	Still too many results, need to refine more
17.9.2019	"Medical genetics" AND software tool AND gene testing AND digital tool	4070	Better, but still too much. Also, the "digital" might be a confusing term.
17.9.2019	"Medical genetics software" AND gene testing	0	Too specific, no results
17.9.2019	**"Medical genetics" AND "software tool" AND gene testing**	476	Much better result, suitable amount to go through.
IEEE Xplore			
Date	Search terms	Result	Comment
20.9.2019	Medical AND genetics AND software	17	Opposite problem to Scholar, too few results. Need to include more options.
20.9.2019	Medical AND genetics AND software OR Medical AND genetics AND web-based	18	Not much better. Could include an alternative word for genetics (genomic).
20.9.2019	**Medical AND genetics AND software OR Medical AND genetics AND web-based OR Medical AND genomics AND software OR Medical AND genomics AND web-based**	31	Not very much, but better. Might be as good as it gets without sacrificing the precision.
PubMed			
Date	Search terms	Result	Comment
20.9.2019	Medical AND genetics AND software	4744	Too many results, need to refine
20.9.2019	**Medical AND genetics AND software OR Medical AND genetics AND web-based OR Medical AND genomics AND software OR Medical AND genomics AND web-based**	157	Tried the same search as in IEEE, seemed to work.

ISI Web of Knowledge			
Date	Search terms	Result	Comment
7.10.2019	Medical AND genetics AND software	62	Quite good but could include the alternative search words also.
7.10.2019	**Medical AND genetics AND software OR Medical AND genetics AND web-based OR Medical AND genomics AND software OR Medical AND genomics AND web-based**	122	Same search as with IEEE Xplore and PubMed, gives nice amount of results.
Scopus			
Date	Search terms	Result	Comment
10.10.2019	Medical AND genetics AND software	613	Too many, try the same search as with some others.
10.10.2019	**Medical AND genetics AND software OR Medical AND genetics AND web-based OR Medical AND genomics AND software OR Medical AND genomics AND web-based**	27	Very manageable result.
ACM Digital Library			
Date	Search terms	Result	Comment
10.10.2019	**Matches all: Medical genetics software**	20	Good result. Did not find option to include alternative words like "Genomic" and "Web-based"

Table 2. Date and search terms used in the test searches of the literature databases. The ones that were used in the actual search are in bold.

Objective was to search for articles published in the range of 2015-2019. The searches were done before 2019 was over (September and October of 2019), thus 2019 is not a full year. Also, for the sources that had an option to restrict the search for the last 5 years also gave some results for articles published in 2014.

To keep the workload and time used on it suitable for Master's thesis, the search was made a little stricter (using "software tool" instead of "software" and adding "Gene testing" to the search terms) and not include some alternative words (like "Genomic" as an alternative to "Genetics") on Google Scholar, which gives results very on a very wide scope. Also, option for including alternatives words ("Genomic" and "Web-based") in one search on ACM Digital Library was not found, and separate searches for them were not done for the same reason.

2.5 Literature trimming and categorizing

After the literature search, I went through the results one by one. I put the references in an Excel file to ease keeping track of the process. Duplicate results were removed. Initial trimming was done based on if the literature type was an article in a scientific journal (books, conference papers and abstracts, posters etc. were eliminated). Then I looked if the article was in English (articles in other languages were eliminated) and if the full article was available, overall or for a University student (e.g. if only abstract was available, the result was eliminated). Finally, I looked at the title and abstract (when needed) to see if the article was about humans (articles dealing with subject other than humans were eliminated).

The articles that survived the initial trimming were looked at more closely. I read the title, abstract and, when needed, other parts (mostly Materials and Methods) to determine if the article was about medical genetics software. Medical genetics software needed to be covered in the article, either as a clearly stated focus of the article or as one of the significant focuses e.g. with comparisons or extensive description of different software. Short description of the software was not enough. In cases with multiple software described or compared, preferably the software were also addressed in the Discussion-portion of the article, but it was not always easy to determine which part would correspond to the Discussion-portion (due to articles having different formats). Thus, some judgement calls were made about if the article was deemed to address software deeply enough. If the article was determined to be about medical genetics software, it was included in the final dataset. Papers about general genetics software were excluded, if applicability to medical genetics was not at least mentioned. Papers about protein function and metabolomics were also excluded, since they are not strictly

about genetics, even if they are related and genetics is often mentioned in them. Included articles are shown in Appendix A.

There were 13 uncertain cases that were also looked at by the supervisor of this Master's thesis, Ville Isomöttönen, to see if we would come to the same conclusion about including or excluding the articles. There was disagreement whether to include the article or not on two cases. They were included and inclusion criteria were clarified on how much of a focus the medical genetics software needed to be in the article for it to be included. Altogether six articles out of the 13 were included.

The articles that made it into the final dataset were looked at still more closely to categorize them based on the research approach, technological focus and journal type and to determine if the data privacy is addressed in the paper. Categorizing the articles into research approach and technological focus groups were determined by reading the abstract and other parts of the text when necessary to determine which category the paper belongs to. Journal type of the article was determined by the name of the journal and, when needed, looking at the description of the journal. Word searches were made to determine whether data privacy is addressed in the paper. Words used in the search were 'confidentiality', 'protection', 'privacy', 'sensitive' and 'secure'. If there was a mention or wider handling of this issue, the article was categorized as addressing data privacy.

3 Results

There were 833 results for the search of articles from different sources. Included in the systematic map were 126 articles, which is about 15.13% of all search results. In Table 3, the search results and number and percentage of included articles are shown for each source of literature, as well as the date the search was made. Figure 4 shows the mean number of articles each year from each source.

Source	**Search Date**	**Search results**	**Included**	**Percentage**
Google Scholar	17.9.2019	476	35	7.35%
IEEE Xplore	20.9.2019	31	12	38.71%
PubMed	20.9.2019	157	53	33.76%
ISI Web of Knowledge	7.10.2019	122	20	16.39%
Scopus	10.10.2019	27	2	7.41%
ACM Digital Library	10.10.2019	20	4	20.00%

Table 3. Sources, date and results of literary search, as well as the number and percentage of articles included in the systematic map.

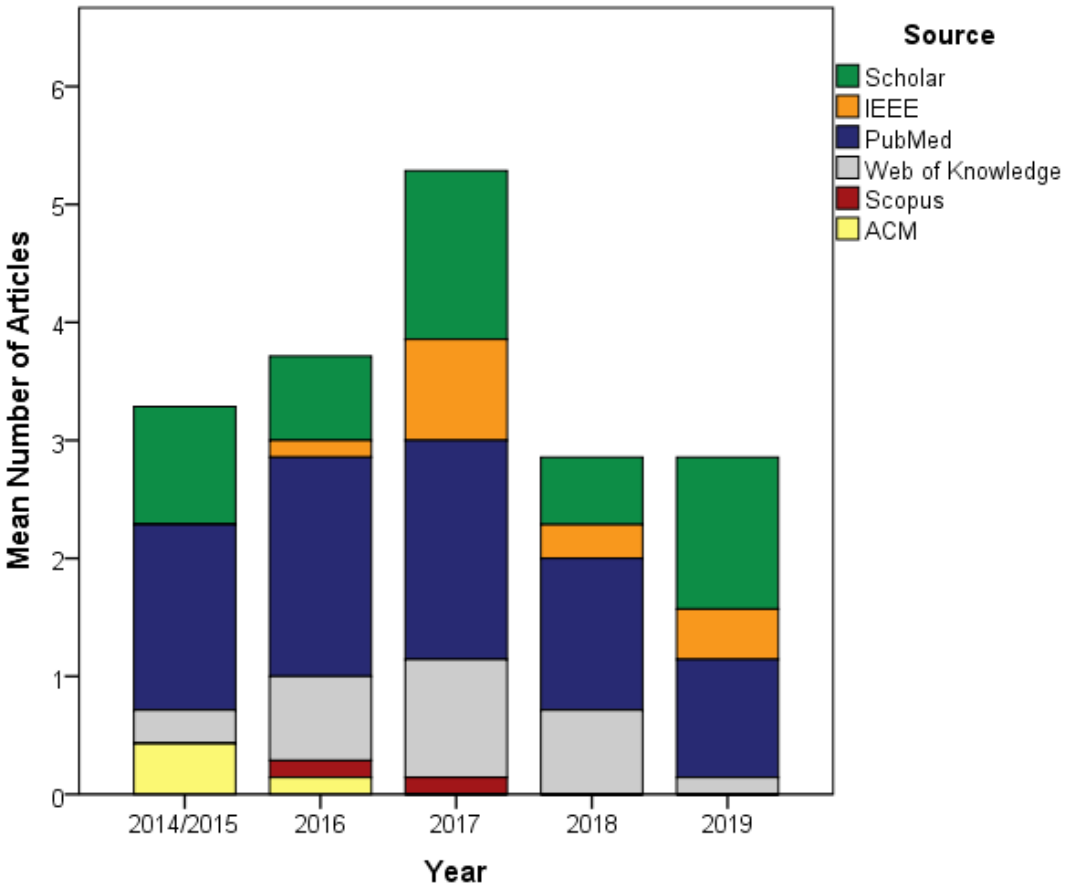

Figure 4. Mean number of articles each year shown from each source.

Duplicates from other sources were left out of the systematic map in the order that the results were analyzed, e.g. if there was the same article in IEEE Xplore and in PubMed, it was left out of included articles in PubMed but kept in IEEE Xplore. Consequently, some sources would have had more included articles than is in Table 3 if duplicates had not been eliminated (Pubmed 54 instead of 53, ISI Web of Knowlegde 25 instead of 20, Scopus 9 instead of 2).

Range of the publishing years was set to 2015-2019, but since some of the sources only had the option of “last five years”, there are some articles from 2014. They are only from the last few months of 2014, so they are combined with articles from 2015.

3.1 Research type

Articles were categorized based on the research approach that was used in the study. Vast majority of the included articles are of type Validation (90 out of 126). Number of articles in each group is shown in Table 5 and Figure 5 shows the different types of articles each year.

Research Type	Number of Articles
Validation	90
Evaluation	9
Philosophical	17
Solution Proposal	8
Experience	2
Total	126

Table 4. Number of articles in each Research Type –group.

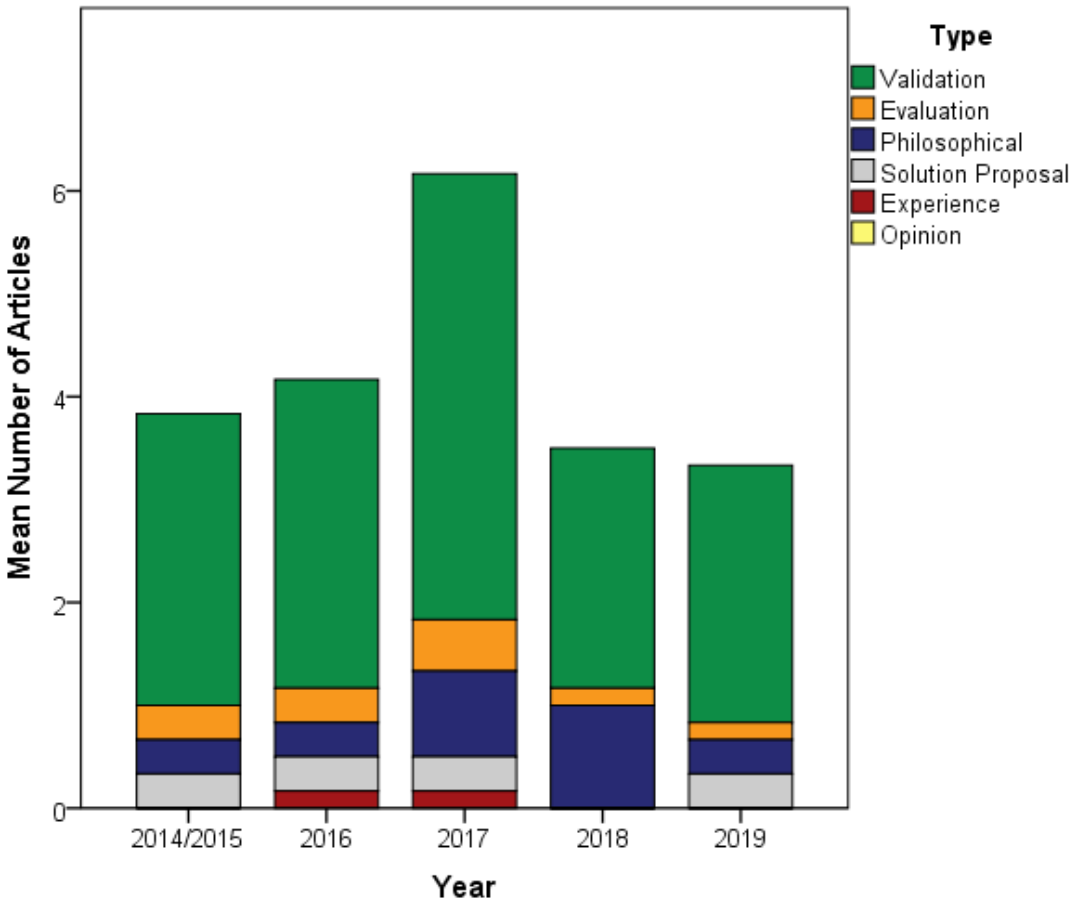

Figure 5. Mean number of articles each year in each research type.

3.2 Technological focus and privacy of data

Articles were also categorized based on the technological focus of the article, i.e. if they are about data storage, data analysis or preprocessing, or about interpreting or visualizing data. Often clear cut categorization was not possible, since article would deal with multiple technological focuses or the software presented had different functions that fell into different categories. Thus, some articles were categorized into combinations of the categories.

Largest group is Analysis (52 out of 126). Number of articles in each group are shown in Table 5 and Figure 6 shows the different groups in each year.

Technological focus	Number of Articles
Storage	8
Analysis	52
Interpretation	12
Storage/Analysis	18
Storage/Interpretation	2
Analysis/Interpretation	11
All	23
Total	126

Table 5. Number of articles in each Technological focus –group.

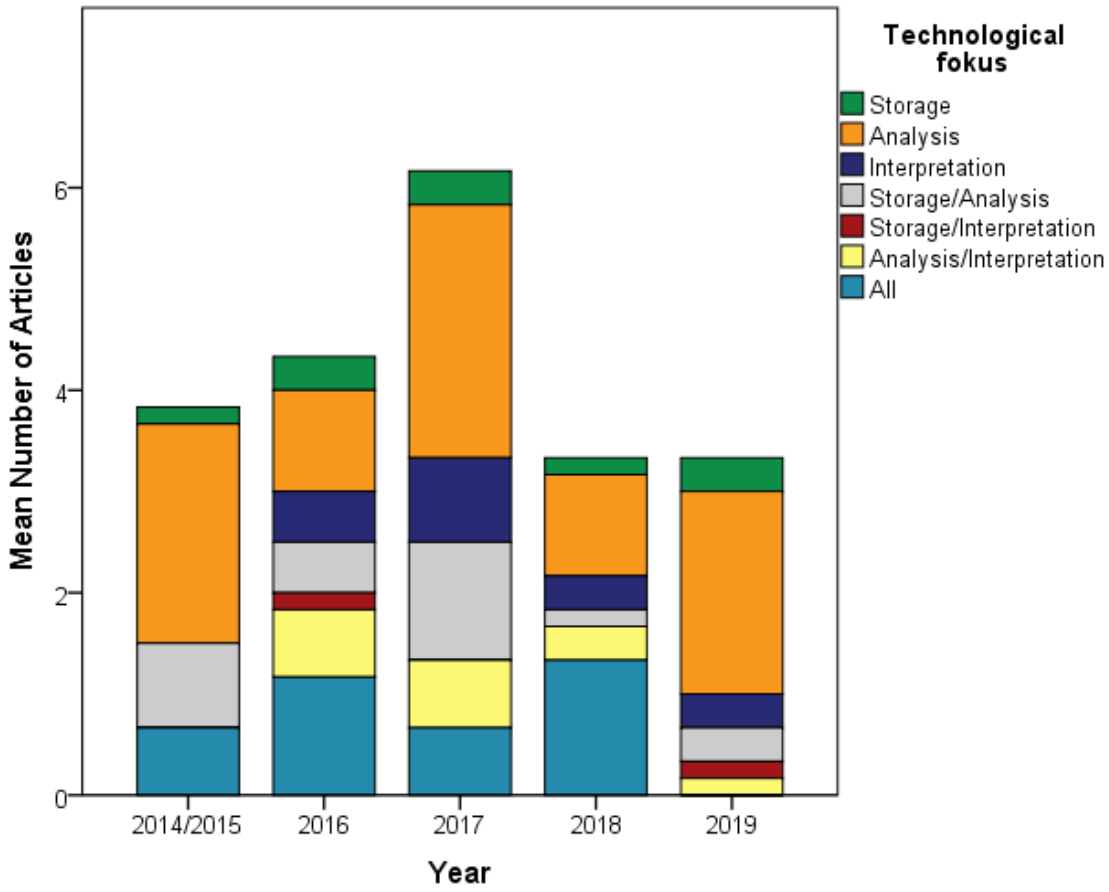

Figure 6. Mean number of articles each year in each Technological focus –group.

One of the research questions was if privacy of data is addressed in the articles. Number of articles addressing data privacy is 35 (out of 126). Privacy of data is addressed most in articles where the Technological focus is combination of all the groups (Storage, Analysis and Interpretation). Table 6 shows the numbers of articles addressing data privacy in each Technological focus groups overall and the percentage of those papers in each group out of all the papers in the group. Figure 7 shows the number of articles addressing data privacy in each of those groups in each year.

Technological focus	**Number of papers privacy is addressed in**	**Percentage of papers in the group**
Storage	2	25 %
Analysis	5	9.62 %
Interpretation	5	41.67 %
Storage/Analysis	5	27.78 %
Storage/Interpretation	1	50 %
Analysis/Interpretation	3	27.27 %
All	14	60.87 %
Total	35	27.78 %

Table 6. Number of articles addressing data privacy in each Technological focus –group and the percentage of those papers in each group out of all the papers in the group.

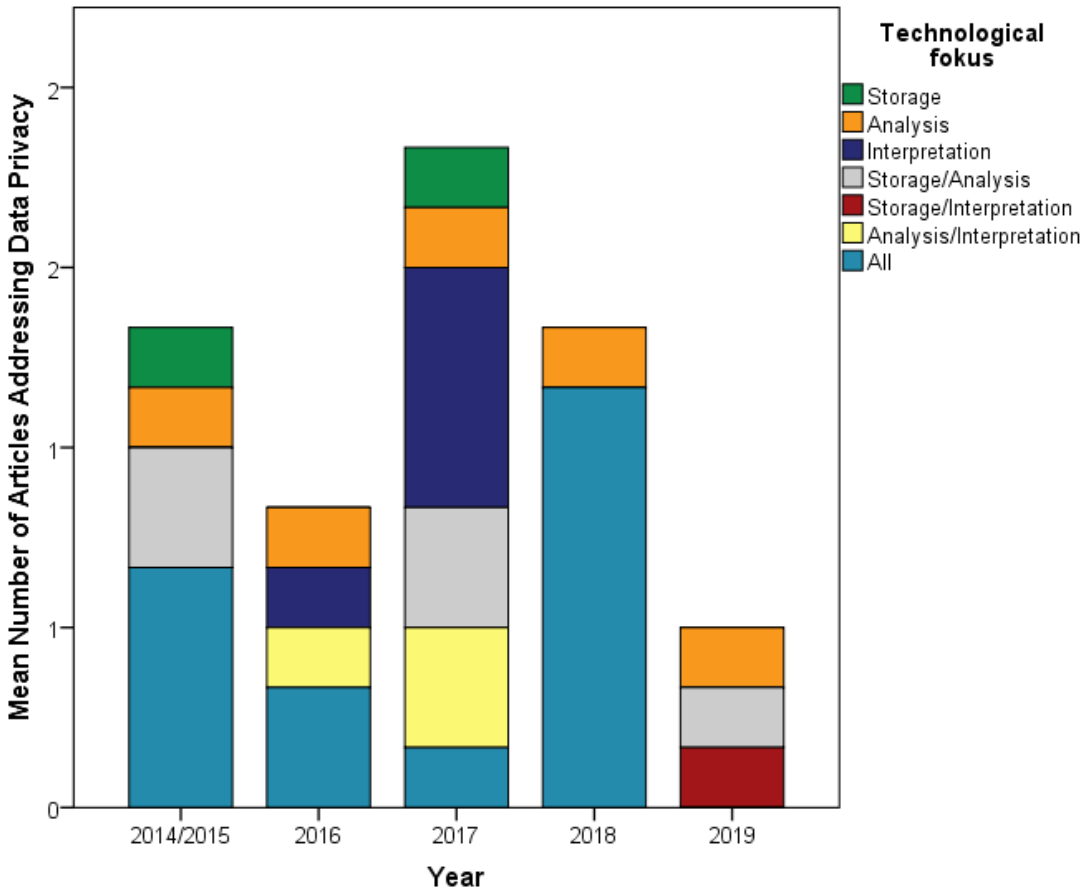

Figure 7. Mean number of articles addressing data privacy each year in each Technological focus -groups

3.3 Journal types

The types of journals the articles were published in was looked into also. Journal types were determined while looking at the included articles and what kinds of groups could be formed from the types of journals the papers were published in based on the focus of the journal. Each paper was categorized based on the title of the journal it was published in, and in unclear cases the description of the journal was checked. Journal types and what kinds of journals are in each type are presented in Table 7.

Journal Type	Kinds of journals included
Bioinformatics	Journals with Bioinformatics on their name and Biodata Mining.
Genetics/DNA	Journals with words like Genome, Genomics or Genetics on their name (without something referring to medical), Nucleic Acid Research and Human Mutation.
Medical	Journals with words like Medicine, Clinical, Epidemiology, Hepatology, Pediatrics or Cancer in the name.
Computer Science	Elife, Gigascience, IEEE Access and Computer Networks
General Natural Sciences	Journals without specific focus but about Natural Sciences, e.g. Scientific reports, Plos One, Nature and Methods.
Patent	Patent applications.
Cell/Yeast/Nano	Journals about cells, yeast or nanotechnology are combined together since there were only few.

Table 7. Different journal types and what kinds of journals are in them.

The largest group was journal type Medical (39 articles) and Genetics/DNA was a close second (36 articles). Table 8 shows the number of articles in each group and Figure 8 shows the mean number of articles in each group for each year.

Journal Type	Number of Articles
Bioinformatics	25
Genetics/DNA	36
Medical	39
Computer Science	6
General Natural Sciences	15
Patent	1
Cell/Yeast/Nano	4
Total	126

Table 8. Number of articles in each Journal Type –group.

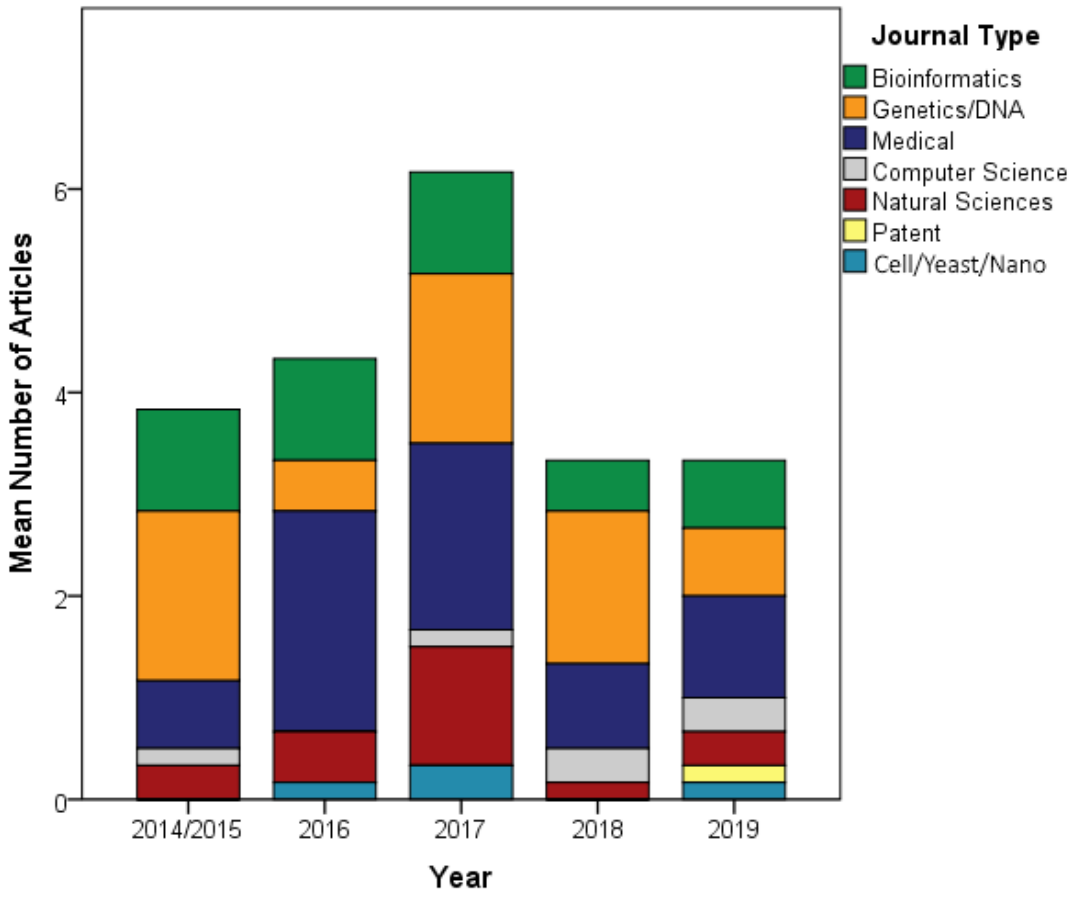

Figure 8. Mean number of articles each year in each Journal Type.

4 Discussion

Using software in medical genetics is becoming more and more important as commercial gene tests and whole-genome sequencing to aid health care are becoming more common (Evans et al. 2016, McGrath and Ghersi 2016). The purpose of this study was to systematically map the literature on medical genetics software to answer four research questions:

1. What types of research approaches there are in the papers?
2. What is the technological focus (Storage, Analysis or Interpretation) of the papers?
3. Is data privacy addressed in the papers?
4. What types of journals the papers are published in?

Search was done on six different sources and results were trimmed to include only the relevant articles published in peer-reviewed journals in 2015-2019. Included papers were categorized to answer the research questions. Results are discussed in the following subchapters.

4.1 Research type

The categorization based on the research approach used shows that Validation is the most common research type in this study. This is probably partly due to what search terms were used and how the inclusion and exclusion criteria were defined. Search was for medical genetics software and papers were only included if that was at least one of the focuses in the paper. It would stand to reason that most of the papers would then be about introducing software that can be used in medical genetics. One of the issues was whether to categorize some of the papers as Validation or Evaluation. Some papers that introduced new software included only small amount of testing of the software whereas some had very extensive testing done either with simulated or real data, or both. To determine what amount of testing would warrant categorizing the paper as Evaluation was difficult, so it was decided to use a very simple and clear cut criteria: if paper was by the developers of the software introducing new software, it was categorized as Validation and if the paper was by other people than the developers of the software, e.g. comparing the software to other software, it was categorized

as Evaluation. This criteria for categorizing is not without problems, since some of the papers categorized as Validation could be better suited to Evaluation.

The second largest group is Philosophical papers. These are the kinds of papers that focus on bringing together information on some subject related to medical genetic software. Some are generally about medical genetics, and some are about some specific system or group of software types (e.g. databases for medical genetics). It could be that these kinds of papers will get more common as genetics starts to be used more and more in health care. In my data, there is some increase in number of papers of this group in more recent years (2017 and 2018) compared to earlier years (2014/2015 and 2016). The exception is 2019 with fewer papers in this group, but that could be because the results do not cover the last few months of 2019 (searches were done in September and October of 2019).

There are some Evaluation papers, but the relatively low number of them could be due to the categorization criteria, as explained above. Some Solution Proposal papers were also found with the search. They are the kinds of papers which outline an idea for a system or software for medical genetics, even with some execution of the idea in some of them, but no testing of system or software included. Relatively low number of these could be due to inclusion criteria only covering articles in peer-reviewed journals and excluding e.g. conference papers, which might have more of Solution Proposal papers. Very low number of Experience papers and no Opinion papers might also be explained by similar reasons.

4.2 Technological focus and privacy of data

Software tools can ease the challenges in using genetics in health care, which include storing and sharing vast amounts of data, analyzing complex genetic data to find relevant information and helping medical care professionals interpret and visualize genetic data (McGrath and Ghersi 2016, Milicchio et al. 2016, Reali et al. 2018, Zhang et al. 2018). In my study, most of the papers deal with preprocessing and analysis of the genetic data. Many different things in genetics are of interest in health care, e.g. small mutations, larger mutation, repeat or copy number variation and methylation of DNA (Read 2017). There are also many dif-

ferent techniques to generating genetic data, from specific gene tests to whole-genome sequencing, and from looking at the structure of chromosomes to measuring gene expression (Hedenfalk et al 2001, Stranger et al. 2005, Klonowska et al. 2015, Reali et al. 2018). It would stand to reason that there is a need for many different software to help with all the different kinds of data.

Relatively low number of papers on storing the data could be explained with many of the databases being established before the range (2015-2019) of this study so that not that many papers are written about them anymore. One reason could also be a shift from merely storing and sharing the data to including tools that enable using the data more efficiently. In my study, the number of papers dealing with not just storage, but storage combined with analysis, interpretation or both is quite high (45), supporting this conclusion.

Relatively low number of papers on interpretation and visualization of genetic data could be due to the need for these kinds of software only recently becoming more into focus. Genetics is just starting to be more common in health care and which the rise of commercial gene tests and whole-genome sequencing, there is starting to be need for regular health care professionals to be able to interpret genetic data (Tinkle and Cheek 2002, McGrath and Ghersi 2016, Zhang et al. 2018). On the other hand, part of the reason for relatively low number of papers on interpreting genetic data might be that there are not so many solutions for just interpreting the data. Interpretation might often be connected to either storing the data (e.g. comparing own data with data from a database and interpreting the meaning of it) or analyzing the data (e.g. data is analyzed with a software and there is a component in the software to help with the interpretation of the results). In my study the number of papers that are dealing with interpretation, but also with storage, analysis or both is relatively high (36).

Only a little over a fourth of all the included papers address the issue of data privacy. Given that data privacy is one of the major concerns in medical genetics (Fuller et al. 1999, Reali et al. 2018, Thorogood et al. 2018), this might be a little surprising. On the other hand, many applications, especially for analyzing the data, are used on the researchers computer (i.e. are not web-based or use cloud services etc.) and data is not shared with others. Data privacy might not be as much of a concern in those kinds of software, or it might be overlooked more

easily. This is reflected in my results with Analysis group having the lowest percentage of papers addressing data privacy. Although data privacy is an issue that should be paid attention to also with desktop software, it becomes even more pertinent e.g. with databases or web-based applications in which other people could have access to the data. In my study, percentage of papers addressing data privacy is higher in Interpretation-group than in Storage group. On the other hand, it is highest in the group that deals with all three technological focuses, Storage, Analysis and Interpretation, and is quite high in the other combination-groups also. It could be that software for interpretation are more often web-based, and more focus is given to data privacy when software are comprehensive, covering multiple technical focuses. Those kinds of software could also more often have some web-based components in them.

Papers addressing data privacy in each group changes a lot in each year, but there does not seem to be any trend to it that would enable making conclusion about it.

4.3 Journal types

Most of the papers were published in medical journal and in genetics/DNA journals. Quite many were also in bioinformatics journals. Relatively low number of papers were published in computer science -type journals. It could be that when articles on software for medical genetics is rather published in fields related to medical genetics instead of e.g. in more general computer science journals. The ones who are using the software are mostly professionals in genetics or health care, that are more likely to read journals in their own field. Software designed specifically for medical genetics is also probably of little interest to computer scientists. Natural Sciences being fourth largest group indicates that there might be some general interest also to this subject, but it is still among the natural sciences. Natural sciences might be generally more interesting to people from different backgrounds in the scientific field. A kind of intersection where medical professionals, geneticists and computer scientists can share interests and present interdisciplinary science.

Only one patent application was in the included papers. This is probably due to Google Scholar being only one of the sources that included an option to search for patent applications

also. The grouping of Cell/Yeast/Nano was more about making a group for the few papers that did not fit the other groups.

Amount of papers in each Journal Type –groups were relatively stable or showed no apparent trend in their changes throughout the years, so no conclusion can be made from that.

4.4 Conclusions

Using genetic data to help with health care is a complex issue with many challenges. The benefits of it, both already existing and potential, are numerous and it could even be called a revolution in the field of health care (McGrath and Ghersi 2016). In my study, the literature seems to reflect the needs of the field in that most of the articles are about preprocessing and analyzing the data, which is arguably the most complex part of using genetic data in health care. On the other hand, the need for software for interpreting and visualizing genetic data is probably going to rise as commercial gene tests start to become more common and more often integrated into the precise care of individuals. In looking at the articles that are included, very few of them are about software specifically for commercial gene tests, which is likely going to change in the future. There is also likely to be a need for software that are multifaceted and able to combine different aspects of handling and using genetic data to make it easier to get relevant information out of genetic data fast and efficiently to help with e.g. urgent situations. It is also easier for medical professionals to have to learn one software that combines the different aspects of handling genetic data than several software for each aspect. This is also reflected in the results as the number of articles that address more than one technological focus is quite high.

The risks of private genetic data ending up in the wrong hands is a serious issue with potentially very severe consequences. The low number of articles that address data privacy could be interpreted that this issue is not taken seriously enough in the field of medical genetics software. Low number of articles addressing the issue could partly be explained by the focus being on software for analysis, which are often not web- or cloud-based. However, security of data this personal should be addressed on those kinds of software also. Security risks are not solely the problem of web-based software. The computers are often connected to the

internet even if the software is not web-based, opening them to attacks. Data should also be anonymized and protected against physical stealing of the data, e.g. the device being stolen. The issue of data privacy will hopefully be addressed more in future literature on the subject.

Conducting this study had some challenges. There could have been more results and thus a better coverage of the literature on medical genetics software, if alternative search terms were used on Google Scholar and ACM Digital Library, as was used on the other sources of literature. Results could have been more versatile if e.g. conference papers were included also. Medical genetics software is quite wide subject and more in-depth systematic mapping might have been achieved with e.g. focusing on some specific types of software. On the other hand, the workload and scope of this study were suitable for a Master's thesis and this study gives a good foundation for expanding it into an academic publication. It could be debated whether all the inclusion criteria and determining the categories were the best possible. I used my own judgement on how to decide between Validation and Evaluation, as well as determining the groups of journal types. Different judgement calls could have been made, resulting in different outcome from this study. However, this study gives a good insight into the field of medical genetics software. It reveals what the focus is in the literature and what type of literature is lacking, giving both information about the situation and possibly ideas about what could be focused more on in the future. It also brings into light that data privacy is not addressed often enough in the literature, which could make professionals in the field pay more attention to this issue in the future.

Bibliography

Abel, E., Horner, S. D., Tyler, D. & Innerarity, S. A. 2005. “The Impact of Genetic Information on Policy and Clinical Practice.” *Policy, Politics & Nursing Practice* 6: 5-14.

Besenbacher, S., Sulem, P., Helgason, A., Helgason, H., Kristjansson. H., Jonasdottir, A., Jonasdottir, A, Magnusson, O. Th., Thorsteinsdottir, U., Masson, G., Kong, A., Gudbjartsson, D. F. & Stefansson, K. 2016. “Multi-nucleotide de novo Mutations in Humans.” *PLoS Genet* 12(11): e1006315.

Brookes, A. J. & Robinson, P. N. 2015. “Human genotype–phenotype databases: aims, challenges and opportunities.” *Nature Reviews Genetics* 16: 702–715

Chang, V. 2018. *“*Data analytics and visualization for inspecting cancers and genes*.” Multimedia tools and applications* 77(14): 17693-17707.

Costello, J. F. & Plass, C. 2001. “Methylation matters.” *Journal of Medical Genetics* 38:285-303.

Evans, J., Wilhelmsen, K., Berg, J., Schmitt, C., Krishnamurthy, A., Fecho, K. & Ahalt, S. 2016. “A new framework and prototype solution for clinical decision support and research in genomics and other data-intensive fields of medicine.” *EGEMS* (Wash DC) 4:1198.

Fredrich, B., Schmöhl, M., Junge, O., Gundlach, S., Ellinghaus, D., Pfeufer, A., Bettecken, T., Siddiqui, R., Franke, A., Wienker, T. F., Hoeppner, M. P. & Krawczak, M. 2019. “VarWatch—A stand-alone software tool for variant matching.” *PLoS ONE* 14(4): e0215618

Fuller, B. P., Ellis Kahn, M. J., Barr, P. A., Biesecker, L., Crowley, E., Garber, J., Mansoura, M. K., Murphy, P., Murray, J., Phillips, J., Rothenberg, K., Rothstein, M., Stopfer, J., Swergold, G., Weber, B., Collins, F. S. & Hudson, K. L. 1999. “Privacy in Genetics Research.” *Science* 285 (5432): 1359-1361.

Gobalan, K. & John, A. 2016. “Applications of bioinformatics in genomics and proteomics.” *Journal of advanced applied scientific research* 3(1): 29-42.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O., Borg, Å. & Trent, J. 2001. "Gene-Expression Profiles in Hereditary Breast Cancer." *New England Journal of Medicine* 344:539-548.

Khandekar, S., Dive, A. & Munde, P. 2013. "Chromosomal abnormalities – a review." *Central India Journal of Dental Sciences* 4: 35-40.

Klonowska, K., Czubak, K., Wojciechowska, M., Handschuh, L., Zmienko, A., Figlerowicz, M., Dams-Kozlowska, H. & Kozlowski, P. 2016. "Oncogenomic portals for the visualization and analysis of genome-wide cancer data." *Oncotarget*, 7(1), 176–192.

McGrath, S. & Ghersi, D. 2016. "Building towards precision medicine: empowering medical professionals for the next revolution." *BMC Med Genomics* 9:23.

Miller, R., Khromykh, A., Babcock, H., Jenevein, C. & Solomon, B. D. 2017. "Putting the Pieces Together: Clinically Relevant Genetic and Genomic Resources for Hospitalists and Neonatologists." *Hosp Pediatr* 7(2):108-114.

Milicchio, F., Rose, R., Bian, J., Min, J. & Prosperi, M. 2016. "Visual programming for next-generation sequencing data analytics." *BioData Mining* 9:16.

Pasternak, J. J. 2005. *An Introduction to Human Molecular Genetics : Mechanisms of Inherited Diseases.* 2nd ed. Hoboken, N.J.: Wiley-Liss.

Petersen, K., Feldt, R., Mujtaba, S. & Mattsson, M. 2008. "Systematic Mapping Studies in Software Engineering." (Italy), EASE'08 68–77.

Petersen, K., Vakkalanka, S. & Kuzniarz, L. 2015. "Guidelines for conducting systematic mapping studies in software engineering: An update." *Information and Software Technology* 64:1–18.

Rabbani, B., Tekin, M. & Mahdieh, N. 2014. "The promise of whole-exome sequencing in medical genetics." *Journal of Human Genetics* 59: 5-15.

Read, C.Y. 2017. “Primer in Genetics and Genomics, Article 3-Explaining Human Diversity: The Role of DNA.” *Biol Res Nurs* 19(3):350-356.

Reali, G., Femminella, M., Nunzi, E. & Valocchi, D. 2018. “Genomics as a service: A joint computing and networking perspective.” *Computer Networks* 145: 27–51.

Shirts B., Salama, J., Aronson, S., Chung, W., Gray, S., Hindorff, L., Jarvik, G., Plon, S., Stoffel, E., Tarczy-Hornoch, P., Van Allen, E., Weck, K., Chute, C., Freimuth, R., Grunmeier, R., Hartzler, A., Li, R., Peissig, P., Peterson, J., Rasmussen, L., Starren, J., Williams, M. & Overby, C. 2015. “CSER and eMERGE: current and potential state of the display of genetic information in the electronic health record.” *J Am Med Inform Assoc* 22:1231 - 42.

Singh, S., Murphy, B. & O'Reilly, R. 2003. “Involvement of gene–diet/drug interaction in DNA methylation and its contribution to complex diseases: from cancer to schizophrenia.” *Clinical Genetics* 64: 451-460.

Stranger, B. E., Forrest, M. S., Clark, A. G., Minichiello, M. J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S. E., Tavaré, S., Deloukas, P. & Dermitzakis, E. T. 2005. “Genome-Wide Associations of Gene Expression Variation in Humans.” *PLoS Genet* 1(6): e78.

Thorogood, A., Touré, S. B., Ordish, J., Hall, A. & Knoppers, B. 2018. “Genetic database software as medical devices.” *Human Mutation* 39(11): 1702–1712.

Tinkle, M. B. & Cheek, D. J. 2002. “Human genomics: challenges and opportunities.” *J Obstet Gynecol Neonatal Nurs* 31(2):178-87.

Wieringa, R., Maiden, N., Mead, N. & Rolland, C. 2006. “Requirements engineering paper classification and evaluation criteria: a proposal and a discussion”. *Requirements Engineering* 11 (1): 102–107.

Wikimedia Commons, search made on Nov 19, 2019. Original picture from www.genome.gov, Public Domain (reference for Figure 1).

Wikimedia Commons, search made on Nov. 19, 2019. By Leonid 2 - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=28463700 (reference for Figure 2).

Zhang, W., Zhang, H., Yang, H., Li, M., Xie, Z. & Li, W. 2018. "Computational resources associating diseases with genotypes, phenotypes and exposures." *Briefings in Bioinformatics* bby071.

Appendices

A Included Articles and their categories

Explanation for the abbreviations

RT = Research Type (V = Validation, E = Evaluation, P = Philosophical, S = Solution Proposal, Ex = Experience)

TF = Technological Focus (S = Storage, A = Analysis, I = Interpretation, SA = Storage/Analysis, SI = Storage/Interpretation, AI = Analysis/Interpretation, SAI = Storage/Analysis/Interpretation)

DP = Data Privacy (is it addressed, Yes = Y, No = N)

JT = Journal Type (B = Bioinformatics, G = Genetics/DNA, M = Medical, C = Computer science, N = Natural Sciences, V = Various i.e. Cell/Yeast/Nano)

Year	Title	Author	RT	TF	DP	JT
2015	Human genotype–phenotype databases: aims, challenges and opportunities.	Brookes AJ & Robinson PN	P	S	Y	G
2016	Functional assays provide a robust tool for the clinical annotation of genetic variants of uncertain significance.	Woods NT, Baskin R, Golubeva V et al.	E	A	N	M
2015	NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data.	Korneliussen, Sand T & Moltke I	V	A	N	B
2016	Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences.	Twist, Greyson P, Gaedigk A et al.	E	AI	N	M

2016	A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics.	James, Regis A, Ian M et al.	V	SAI	N	M
2015	An integrative approach to predicting the functional effects of non-coding and coding sequence variation.	Shihab, Hashem A, Rogers MF et al.	V	A	N	B
2017	Detection of long repeat expansions from PCR-free whole-genome sequence data.	Dolzhenko E, van Vugt JJFA, Shaw RJ et al.	V	A	N	G
2018	Assessment of the incorporation of CNV surveillance into gene panel next-generation sequencing testing for inherited retinal diseases.	Ellingford JM, Horn B, Campbell C et al.	E	A	N	M
2015	SPS: A Simulation Tool for Calculating Power of Set-Based Genetic Association Tests.	Li J, Sham PC, Song Y et al.	S	A	N	M
2018	RD-Connect, NeurOmics and EURenOmics: collaborative European initiative for rare diseases.	Lochmüller H, Badowska DM, Thompson R et al.	P	SAI	Y	G
2017	Genetic identification of a common collagen disease in Puerto Ricans via identity-by-descent mapping in a health system.	Belbin, Morven G, Odgis J et al.	E	S	N	C
2015	Mosaic structural variation in children with developmental disorders.	King DA., Jones WD, Crow YJ et al.	E	A	N	G
2017	VCF. Filter: interactive prioritization of disease-linked genetic variants from sequencing data.	Müller H, Jimenez-Heredia R, Krolo A et al.	V	AI	Y	G
2017	A practical guide to filtering and prioritizing genetic variants.	Dashti S, Jalali M & Gamieldien J	P	SA	N	N

2017	mirVAFC: A web server for prioritizations of pathogenic sequence variants from exome sequencing data via classifications.	Li Z, Liu Z, Jiang Y et al.	E	AI	N	G
2017	Imbalance-aware machine learning for predicting rare and common disease-associated non-coding variants.	Schubach M, Re M, Robinson PN et al.	V	A	N	N
2017	CRIMEtoYHU: a new web tool to develop yeast-based functional assays for characterizing cancer-associated missense variants.	Mercatanti A, Lodovichi S, Cervelli T et al.	V	SA	N	V
2017	pyAmpli: an amplicon-based variant filter pipeline for targeted resequencing data.	Beyens M, Boeckx N, Van Camp G et al.	V	A	N	B
2016	A computer-assisted method for pathogenicity assessment and genetic reporting of variants stored in the Australian Inherited Retinal Disease Register.	Huynh E, De Roach J, McLaren T et al.	V	SA	N	M
2017	An evaluation of copy number variation detection tools for cancer using whole exome sequencing data.	Zare F, Dow M, Monteleone N et al.	E	A	N	B
2018	Computational resources associating diseases with genotypes, phenotypes and exposures.	Zhang W, Zhang H, Yang H, et al.	P	SAI	Y	B
2015	Challenges in exome analysis by LifeScope and its alternative computational pipelines.	Pranckevičiene E, Rančelis T, Pranculis A et al.	E	A	N	N
2019	VarWatch—A stand-alone software tool for variant matching.	Fredrich B, Schmöhl M, Junge O et al.	P	SA	Y	N
2019	Pharmacogenomics Clinical Annotation Tool (Pharm CAT).	Sangkuhl K, Whirl-Carrillo M, Whaley RM et al.	V	A	N	M

2019	Variant Interpretation for Cancer (VIC): a computational tool for assessing clinical impacts of somatic variants.	He MM, Li Q, Yan M et al.	V	SI	Y	M
2015	CNV-ROC: A cost effective, computer-aided analytical performance evaluator of chromosomal microarrays.	Goodman CW, Major HJ, Walls WD et al.	V	A	N	M
2018	Bioinformatics in Clinical Genomic Sequencing.	Lebo MS, Hao L, Lin C et al.	P	SAI	Y	M
2019	Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings.	Hwang K, Lee I, Li H et al.	P	A	N	N
2019	Bioinformatics-Based Identification of Expanded Repeats: A Non-reference Intronic Pentamer Expansion in RFC1 Causes CANVAS.	Rafehi H, Szmulewicz DJ, Bennett MF et al.	E	A	N	G
2019	InTAD: chromosome conformation guided analysis of enhancer target genes.	Okonechnikov K, Erkek S, Korbel JO et al.	V	A	N	B
2016	A systematic comparison of copy number alterations in four types of female cancer.	Kaveh F, Baumbusch LA, Nebdal D et al.	Ex	A	N	M
2017	XCAVATOR: accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments.	Magi A, Pippucci T & Sidore C	V	A	N	G
2019	CHASMplus reveals the scope of somatic missense mutations driving human cancers.	Tokheim C & Karchin R	V	A	N	V
2019	Systems and methods for predicting genetic diseases.	Zhang S, Li J & Snyder MP	S	A	N	P
2019	Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding.	Hanghøj K, Moltke I, Alstrup P et al.	V	A	N	C

2019	Developing a Multi-Dose Computational Model for Drug-Induced Hepatotoxicity Prediction Based on Toxicogenomics Data.	Su R, Wu H, Xu B et al.	V	A	N	B
2019	SAFETY: Secure gwAs in Federated Environment through a hYbrid Solution.	Sadat MN, Al Aziz MM, Mohammed N et al.	V	A	Y	B
2017	IntSIM: An Integrated Simulator of Next-Generation Sequencing Data.	Yuan X, Zhang J & Yang L	V	A	N	M
2018	Prediction of Drug-Disease Associations for Drug Repositioning Through Drug-miRNA-Disease Heterogeneous Network.	Chen H & Zhang Z	V	A	N	C
2016	Integration and Querying of Genomic and Proteomic Semantic Annotations for Biomedical Knowledge Extraction.	Masseroli M, Canakoglu A & Ceri S	V	SAI	N	B
2019	One Size Does Not Fit All: Querying Web Polystores.	Khan Y, Zimmermann A, Jha A et al.	V	SA	N	C
2017	Towards Unsupervised Gene Selection: A Matrix Factorization Framework.	Li J & Wang F	V	A	N	B
2017	Bosco: Boosting Corrections for Genome-Wide Association Studies With Imbalanced Samples.	Bao F, Deng Y, Zhao Y et al.	V	A	N	V
2018	PerPAS: Topology-Based Single Sample Pathway Analysis Method," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 3, pp. 1022-1027, 1 May-June 2018.	Liu C, Lehtonen R & Hautaniemi S	V	SA	N	B
2017	"MIGS-GPU: Microarray Image Gridding and Segmentation on the GPU.	Katsigiannis S, Zacharia E & Maroulis D	V	A	N	M

2017	D-Map: Random Walking on Gene Network Inference Maps Towards differential Avenue Discovery.	Athanasiadis E, Bourdakou M & Spyrou G	V	A	N	B
2017	Prediction and Validation of Disease Genes Using HeteSim Scores	Zeng X, Liao Y, Liu Y et al.	V	SA	N	B
2019	Pediatric Cancer Variant Pathogenicity Information Exchange (PeCanPIE): a cloud-based platform for curating and classifying germline variants.	Edmonson MN, Patel AN, Hedges DJ et al.	V	A	N	G
2019	7: Comprehensive characterization of a Canadian cohort of von Hippel-Lindaudisease patients.	Salama Y, Albanyan S, Szybowska M et al.	V	I	N	M
2019	Dynamics and predicted drug response of a gene network linking dedifferentiation with beta-catenin dysfunction in hepatocellular carcinoma.	Gérard C, Di-Luoffo M, Gonay L et al.	V	A	N	M
2019	VGSC2: Second generation vector graph toolkit of genome synteny and collinearity.	Xu Y, Wang Q, Tanon Reyes L et al.	S	A	N	G
2019	DEBrowser: interactivedifferential expression analysis and visualization tool for count data.	Kucukural A, Yukselen O, Ozata DM et al.	V	S	N	G
2018	CellMinerCDB for Integrative Cross-Database Genomics and Pharmacogenomics Analyses of Cancer Cell Lines.	Rajapakse VN, Luna A, Yamade M et al.	V	A	N	N
2019	ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles.	Silva TC, Coetzee SG, Gull N et al.	V	I	N	B
2018	HiGlass: web-based visual exploration and analysis of genome interaction maps.	Kerpedjiev P, Abdennur N, Lekschas F et al.	V	AI	N	G

2018	WHAM!: a web-based visualization suite for user-defined analysis of metagenomic shotgun sequencing data.	Devlin JC, Battaglia T, Blaser MJ et al.	V	A	N	G
2018	The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update.	Afgan E, Baker D, Batut B, van den Beek M et al.	P	SAI	Y	G
2018	Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic,transcriptional and epigenomic signatures.	Lee J, Lee AJ, Lee JK et al.	V	A	N	G
2018	ClinVar Miner: Demonstrating utility of aWeb-based tool for viewing and filtering ClinVar data.	Henrie A, Hemphill SE, Ruiz-Schultz N et al.	V	AI	N	G
2018	BRepertoire: a user-friendly web server for analysing antibody repertoire data.	Margreitter C, Lu HC, Townsend C et al.	V	A	N	G
2018	GENEASE: real time bioinformatics tool for multi-omics and disease ontology exploration, analysis and visualization.	Ghandikota S, Hershey GKK & Mersha TB.	V	SAI	N	B
2019	Clonal expansion across the seas as seen through CPLP-TB database: A joint effort in cataloguing Mycobacterium tuberculosis genetic diversity in Portuguese-speaking countries.	Perdigão J, Silva C, Diniz J et al.	V	S	N	M
2018	Human ring chromosome registry for cases in the Chinese population: re-emphasizing Cytogenomic and clinical heterogeneity and reviewing diagnostic and treatment strategies.	Hu Q, Chai H, Shu W et al.	V	S	N	M
2017	Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists.	Zhu X, Wolfgruber TK, Tasato A et al.	V	I	Y	M

2017	The TB Portals: an Open-Access, Web-Based Platform for Global Drug-Resistant-Tuberculosis Data Sharing and Analysis.	Rosenthal A, Gabrielian A, Engle E et al.	V	SA	N	M
2017	geneSurv: An interactive web-based tool for survival analysis in genomics research.	Korkmaz S, Goksuluk D, Zararsiz G et al.	V	A	N	M
2017	India AlleleFinder: a web-based annotation tool for identifying common alleles in next-generation sequencing data of Indian origin.	Zhang JF, James F, Shukla A et al.	V	SA	N	N
2017	Mendel,MD: Auser-friendly open-source web tool for analyzing WES and WGS in the diagnosis of patients with Mendelian disorders.	Cardenas R, Linhares N, Ferreira R et al.	V	AI	N	N
2017	PathOS: a decision support system for reporting high throughput sequencing of cancers in clinical diagnostic laboratories.	Doig KD, Fellowes A, Bell AH et al.	V	SA	Y	M
2017	DeSigN: connecting gene expression with therapeutics for drug repurposing and development.	Lee BK, Tiong KH, Chang JK et al.	V	A	N	G
2017	Putting the Pieces Together: Clinically Relevant Genetic and Genomic Resources for Hospitalists and Neonatologists.	Miller R, Khromykh A, Babcock H et al.	P	SAI	Y	M
2017	ClinGen Resource. ClinGen Pathogenicity Calculator: a configurable system for assessing pathogenicity of genetic variants.	Patel RY, Shah N, Jackson AR et al.	V	A	N	M
2017	The Papillomavirus Episteme: a major update to the papillomavirus sequence database.	Van Doorslaer K, Li Z, Xirasagar S et al.	P	SAI	N	G
2017	The UCSC Genome Browser database: 2017 update.	Tyner C, Barber GP, Casper J et al.	P	SAI	Y	G

2016	eFORGE: A Tool for Identifying Cell Type-Specific Signal in Epigenomic Data.	Breeze CE, Paul DS, van Dongen J et al.	V	SAI	N	V
2017	My46: a Web-based tool for self-guided management of genomic test results in research and clinical settings.	Tabor HK, Jamal SM, Yu JH et al.	S	I	Y	M
2016	Integrating genomic information with protein sequence and 3D atomic level structure at the RCSB protein data bank.	Prlic A, Kalro T, Bhattacharya R et al.	V	SI	N	B
2016	ExSurv: A Web Resource for Prognostic Analyses of Exons Across Human Cancers Using Clinical Transcriptomes.	Hashemikhabir S, Budak G & Janga SC.	V	SAI	N	M
2016	BioVLAB-mCpG-SNP-EXPRESS: A system for multi-level and multi-perspective analysis and exploration of DNA methylation, sequence variation (SNPs), and gene expression from multi-omics data.	Chae H, Lee S, Seo S, Jung D et al.	V	A	Y	N
2016	The Cancer Epidemiology Descriptive Cohort Database: A Tool to Support Population-Based Interdisciplinary Research.	Kennedy AE, Khoury MJ, Ioannidis JP et al.	V	S	N	M
2016	CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets.	Schofield EC, Carver T, Achuthan P et al.	V	A	N	B
2016	MSeqDR: A Centralized Knowledge Repository and Bioinformatics Web Resource to Facilitate Genomic Investigations in Mitochondrial Disease.	Shen L, Diroma MA, Gonzalez M et al.	V	SAI	Y	G
2016	An interactive web-based application for Comprehensive Analysis of RNAi-screen Data.	Dutta B, Azhir A, Merino LH et al.	V	SAI	Y	N

2016	L1000CDS(2): LINCS L1000 characteristic direction signatures search engine.	Duan Q, Reid SP, Clark NR et al.	V	A	N	N
2015	Precise genotyping and recombination detection of Enterovirus. BMC Genomics.	Lin CH, Wang YB, Chen SH et al.	V	A	N	G
2016	HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes.	Forster SC, Browne HP, Kumar N et al.	V	SA	N	G
2015	A Database of Gene Expression Profiles of Korean Cancer Genome.	Kim SK & Chu IS.	V	SA	N	G
2016	Colorectal cancer atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues.	Chisanga D, Keerthikumar S, Pathan M et al.	V	S	N	G
2016	Web-based Gene Pathogenicity Analysis (WGPA): a web platform to interpret gene pathogenicity from personal genome data.	Diaz-Montana JJ, Rackham OJ, Diaz-Diaz N et al.	S	AI	N	B
2016	Oncogenomic portals for the visualization and analysis of genome-wide cancer data.	Klonowska K, Czubak K, Wojciechowska M et al.	P	SAI	N	M
2015	The UK10K project identifies rare variants in health and disease.	UK10K Consortium, Walter K, Min JL et al.	V	SA	N	N
2015	Innovative genomic collaboration using the GENESIS (GEM.app) platform.	Gonzalez M, Falk MJ, Gai X et al.	V	SA	Y	G
2017	BioBankWarden: A web-based system to support translational cancer research by managing clinical and biomaterial data.	Ferretti Y, Miyoshi NSB, Silva WA Jr et al.	V	SA	Y	M

2015	Mitochondrial Disease Sequence Data Resource (MSeqDR): a global grass-roots consortium to facilitate deposition, curation, annotation, and integrated analysis of genomic data for the mitochondrial disease clinical and research communities.	Falk MJ, Shen L, Gonzalez M, Leipzig J et al.	S	SAI	Y	G
2015	VEGAS2: Software for More Flexible Gene-Based Testing.	Mishra A & Macgregor S	V	A	N	G
2015	Beyond protein expression, MOPED goes multi-omics.	Montague E, Janko I, Stanberry L et al.	V	SAI	Y	G
2015	The UCSC Cancer Genomics Browser: update 2015.	Goldman M, Craft B, Swatloski T et al.	P	SAI	Y	G
2014	GWATCH: a web platform for automated gene association discovery analysis.	Svitin A, Malov S, Cherkasov N et al.	V	SAI	Y	C
2014	VariantDB: a flexible annotation and filtering portal for next generation sequencing data.	Vandeweyer G, Van Laer L, Loeys B et al.	V	SA	Y	M
2014	APPEX: analysis platform for the identification of prognostic gene expression signatures in cancer.	Kim SK, Hwan Kim J, Yun SJ et al.	V	SA	N	B
2016	XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits.	Fang H, Knezevic B, Burnha KL et al.	V	I	N	M
2016	SMART precision cancer medicine: a FHIR-based app to provide genomic information at the point of care.	Warner JL, Rioth,MJ, Mandl KD et al.	V	I	Y	M

2015	Scalable and cost-effective NGS genotyping in the cloud.	Souilmi Y, Lancaster AK, Jung J et al.	V	A	N	M
2015	Pinpointing disease genes through phenomic and genomic data fusion.	Jiang R, Wu M & Li L	V	A	N	G
2017	Combining clinical and genomics queries using i2b2-Three methods.	Murphy SN, Avillach P, Bellazzi R et al.	S	SAI	N	N
2017	PLATO software provides analytic framework for investigating complexity beyond genome-wide association studies.	Hall MA, Wallace J, Lucas A et al.	V	A	N	N
2017	Creating a scalable clinical pharmacogenomics service with automated interpretation and medical record result integration - experience from a pediatric tertiary care facility.	Manzi SF, Fusaro VA, Chadwick L et al.	Ex	I	Y	M
2016	Visual programming for next-generation sequencing data analytics.	Milicchio F, Rose R, Bian J et al.	P	AI	Y	B
2017	Data Management for Heterogeneous Genomic Datasets.	Ceri S, Kaitoua A, Masseroli M et al.	V	A	Y	B
2017	Semi-automated cancer genome analysis using high-performance computing.	Crispatzu G, Kulkarni P, Toliat M et al.	V	AI	Y	G
2016	User-centered design of multi-gene sequencing panel reports for clinicians.	Cutting E, Banchero M, Beitelshees AL et al.	V	I	N	M
2018	Genomics as a service: A joint computing and networking perspective.	Reali G, Femminella M, Nunzi E et al.	P	SAI	Y	C

2018	Implementation of a patient-facing genomic test report in the electronic health record using a web-application interface.	Williams MS, Kern MS, Lerch V et al.	V	SAI	Y	M
2016	WHATIF: An open-source desktop application for extraction and management of the incidental findings from next-generation sequencing variant data.	Ye Z, Kadolph C, Strenn R et al.	V	AI	N	M
2018	Genetic database software as medical devices.	Thorogood A, Toure SB, Ordish J et al.	P	SAI	Y	G
2018	hgvs: A Python package for manipulating sequence variants using HGVS nomenclature: 2018 Update.	Wang M, Callenberg K, Dalgleish R et al.	V	A	Y	G
2018	D3Oncoprint: Stand-Alone Software to Visualize and Dynamically Explore Annotated Genomic Mutation Files.	Palmisano A, Zhao Y & Simon RM	V	I	N	M
2017	GARLIC: a bioinformatic toolkit for aetiologically connecting diseases and cell type-specific regulatory maps.	Nikolic M, Papantonis A & Rada-Iglesias A	V	SA	N	G
2019	DMD Open-access Variant Explorer (DOVE): A scalable, open-access, web-based tool to aid in clinical interpretation of genetic variants in the DMD gene.	Bailey M & Miller N	V	I	N	M
2017	MitoSuite: a graphical tool for human mitochondrial genome profiling in massive parallel sequencing.	Ishiya K & Ueda S	V	I	Y	N
2017	GenBank.	Benson DA, Cavanaugh M, Clark K et al.	P	S	Y	G

2016	Gene discovery for Mendelian conditions via social networking: De novo variants in KDM1A cause developmental delay and distinctive facial features.	Chong JX, Yu JH, Lorentzen P et al.	S	SA	Y	M
2016	Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data.	Bonilla-Huerta E, Hernandez-Montiel A, Morales-Caporal R et al.	V	A	N	B
2015	Software suite for gene and protein annotation prediction and similarity search.	Chicco D & Masseroli M	V	A	N	B
2015	Building transcriptional association networks in cytoscape with RegNetC.	Nepomuceno-Chamorro IA, Marquez-Chamorro A & Aguilar-Ruiz JS	V	A	N	B
2015	Heterogeneous cloud framework for big data genome sequencing.	Wang C, Li X, Chen P et al.	V	A	N	B