

Marianna Niemi

**GUIDELINES FOR CLINICIANS AND  
SCIENTISTS TO MAKE MOST OUT OF  
FINNGEN GENOME AND DIGITAL HEALTH  
CARE DATA - THE FINNGEN ANALYST  
HANDBOOK**

Faculty of Medicine and Health  
Technology

Master's thesis in Biomedical  
Technology

April 2022

# ABSTRACT

NIEMI, MARIANNA: Guidelines for clinicians and scientists to make most out of FinnGen genome and digital health care data - The FinnGen Analyst Handbook

Master's Thesis: 43 pages

Tampere University

Study program: Master's Programme in Biomedical Technology

Supervisor: Professor Matti Nykter

Examiners: Professor Matti Nykter, University lecturer, Dr. Tech Juha Kesseli

April 2022

---

The FinnGen Analyst Handbook is an electronic guidebook aiming to provide FinnGen researchers with all the guidelines, knowledge and helpful tips they need when analysing, interpreting, and making discoveries with the FinnGen data. FinnGen Analyst Handbook provides detailed instructions for conducting genome-wide association study (GWAS) and medical register-based analysis aiming to reveal associations between conditions and the genome.

FinnGen, started in 2017, is a public-private research project funded by Business Finland and 13 pharmaceutical companies. FinnGen's host organization is the Institute for Molecular Medicine Finland, FIMM, University of Helsinki. The aim of FinnGen study is to improve human health through genetic research and lead to improvements in diagnostics and new therapeutic targets for treating numerous human diseases.

FinnGen project combines genome data from 500,000 Finnish biobank participants with a longitudinal lifetime spanning health registry data aiming to provide comprehensive data for research of various human diseases. By finding associations between genetic factors and health outcomes FinnGen project aims to provide novel medically and therapeutically relevant insights. Being one of the biggest Biobank projects worldwide FinnGen provides a world-class resource for future research.

This Master's Thesis work was to write documentation for FinnGen Analyst Handbook. This thesis gives a report about the Analyst Handbook and its writing process. In addition, one example of the entire workflow for GWAS using Analyst Handbook instructions, FinnGen custom-made tools, and R coding is provided.

Keywords: Genome-wide association study, GWAS, Biobank, medical registers, FinnGen, population genetics, human conditions genetics

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

# TIIVISTELMÄ

NIEMI, MARIANNA: Työohjeita klinikoille ja tutkijoille FinnGen genomidatan ja digitaalisen terveysrekisteridatan tehokkaaseen hyödyntämiseen – The FinnGen Analyst Handbook

Pro gradu -tutkielma: 43 sivua

Tampereen yliopisto

Tutkinto-ohjelma: Biolääketieteellisen tekniikan Maisteri ohjelma

Ohjaaja: Professori Matti Nykter

Tarkastajat: Professori Matti Nykter, yliopiston lehtori, TkT Juha Kesseli

Huhtikuu 2022

---

FinnGen Analyst Handbook on elektroninen käsikirja, jonka tavoitteena on tarjota FinnGen-tutkijoille työohjeet, tiedot ja hyödylliset vinkit FinnGen-aineiston analysointiin, tulkintaan ja lääketieteellisesti relevanttien geneettisten löydösten tekoon. FinnGen Analyst Handbook tarjoaa yksityiskohtaiset ohjeet genomilaajuisen assosiaatioanalyysin tekoon (GWAS) ja terveystietoihin perustuviin analyysihin, joiden tavoitteena on löytää assosiaatioita genomien ja sairauksien välillä.

Vuonna 2017 alkanutta FinnGen tutkimusprojektia rahoittaa Business Finland ja 13 kansainvälistä lääkeyhtiötä. FinnGen projektin isäntäorganisaatio on Suomen molekyyllilääketieteen instituutti (FIMM), Helsingin yliopisto. FinnGen projektin tavoitteena on edistää ihmisten terveyttä genomitutkimuksen keinoin parantamalla diagnostiikkaa ja paikantamalla genomista kohteita uusien hoitojen kehittämiseksi.

FinnGen projekti yhdistää genomidataa ja kansallisiin rekistereihin perustuvaa terveystietoa 500000 suomalaiselta biopankkinäytteen luovuttajalta. Löytämällä assosiaatioita geneettisten tekijöiden ja terveydentilan välillä FinnGen projekti tähtää uusiin lääketieteellisesti ja hoidollisesti merkittäviin löytöihin. Yhtenä suurimmista biopankkiaineistoja hyödyntävistä projekteista maailmassa, FinnGen-tutkimus tarjoaa maailmanluokan tutkimusresursseja myös tulevaisuuden tutkimukselle.

Tämän Pro Gradu Maisterityön tehtävä oli tuottaa dokumentaatiota FinnGen Analyst Handbook käsikirjaan. Pro graduissa esitetään raportti Analyst Handbook käsikirjasta ja sen kirjoitusprosessista. Lisäksi annetaan yksi esimerkki GWAS-analyysin työvaiheista Analyst Handbook käsikirjan ohjeistusta noudattaen sekä FinnGenin kustomoituja työkaluja, että manuaalisesti R-koodausta käyttäen.

Avainsanat: Genominlaajuinen assosiaatiotutkimus, GWAS, biopankki, biopankit, digitaalinen lääketieteellinen rekisteridata, FinnGen, populaatiogenetiikka, sairauksien genetiikka

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck –ohjelmalla.

# PREFACE

My deepest gratitude goes to FinnGen research project, PI Aarno Palotie, project manager Mervi Aavikko, Mary Pat Reeve, Mari Kaunisto, and Risto Kajanne. I feel privileged for the opportunity to work for the FinnGen project. I thank the FinnGen Documentation team, PO Susanna Lemmelä, FinnGen e-Science team, team leader Timo Sipilä, and FinnGen Trajectory team, team leader Mary Pat Reeve for great leadership, creating a good team spirit, and for the opportunity to work in your teams. My teammates Javier, Harri, Elina, Ghazal, Vincent, Emma, Rigbe, Alexander, Sanni, Vishal, Kumar, and Anastasia it has been wonderful to work with you with documentation in the Documentation team, tool development in the Trajectory team, and data manager and helpdesk tasks in the e-science team. I thank all FinnGen researchers who have contributed to writing the FinnGen Analyst Handbook.

I thank my home organization Research services, TAYS, Tampere University Hospital, Director of Research Tarja Laitinen, and head of data Leena Hakkarainen for great leadership, and my teammates Sampo, Petri, Nita, Toni, Mika, and Tiina for co-operation.

I thank my supervisor Professor Matti Nykter, Computational biology, Faculty of medicine and health technology, Tampere University, for supervising and reviewing my master thesis in bioinformatics and for the opportunity to join his team as a post-Doctoral researcher. I thank Juha Kesseli for lessons in bioinformatics and for reviewing this thesis. I thank all my teammates in Nykter Lab for being good colleagues and excellent fellow students.

I thank my family for believing in me and for their patience with my everlasting eagerness to study and learn. My husband Lauri and our two wonderful sons Aapo and Vilho. Thank you Ritva, Pekka, Olli, Susanna, Aira, and Eero for your support.

Tampere, 25 April 2022

Marianna Niemi

# CONTENTS

<b>1 INTRODUCTION .....</b>	<b>1</b>
<b>2 BACKGROUND .....</b>	<b>2</b>
2.1 FinnGen Analyst Handbook	2
2.2 FinnGen project	4
2.3 Description of Analyst Handbook users	5
<b>3 MATERIALS AND METHODS .....</b>	<b>7</b>
3.1 The growing project needed detailed documentation	7
3.2 Teams who wrote the FinnGen Analyst Handbook	8
3.3 Technology	8
3.4 User interviews	10
3.5 History of the FinnGen Analyst Handbook - from concept to implementation	11
3.6 The version history of the Analyst Handbook	13
3.7 Publishing of the FinnGen Analyst Handbook	15
<b>4 RESULTS.....</b>	<b>16</b>
4.1 Statistics of the FinnGen Analyst Handbook usage	16
4.2 Example on how to conduct a custom GWAS in FinnGen Sandbox using instructions of the Analyst Handbook	17
4.2.1 Example workflow and cohorts	17
4.2.2 Example on how to create GWAS analysis with FinnGen custom tools	18
4.2.3 Example how to build patient cohorts in R and conduct GWAS from the command line	25
4.2.4 Viewing GWAS results with FinnGen PheWeb tool	32
<b>5 CONCLUSIONS .....</b>	<b>34</b>
<b>6 REFERENCES .....</b>	<b>35</b>
<b>7 APPENDIX.....</b>	<b>37</b>

# LIST OF SYMBOLS AND ABBREVIATIONS

ATC	Coding system for drugs
CLI	command line interface
Custom GWAS CLI	a FinnGen tool for custom GWAS with command line interface
Custom GWAS GUI	a FinnGen tool for custom GWAS with graphical user interface
DF	Data Freeze
DRC	Descendant record counts
FHRB	Hematological Biobank
GDrugWas	GWAS for patient cohorts divided based on drug usage rather than a condition
GUI	graphical user interface
GWAS	genome-wide association study
ID	Identification number
Kela	the Social Insurance Institution of Finland
PheWAS	phenome-wide association study
PheWeb	a FinnGen tool to browse GWAS results made from FinnGen endpoints online. Linking to Risteys provided for each endpoint.
Python	programming language Python
R	programming language R
R06A	start of ATC codes for antihistamines
RC	Record counts
Risteys	a FinnGen tool to browse FinnGen endpoints online
THL	Finnish institute for health and welfare
UKBB	United Kingdom Biobank

# 1 INTRODUCTION

The FinnGen Analyst Handbook [1] is an electronic guidebook aiming to provide FinnGen project researchers with all the guidelines, knowledge and helpful tips they need when analysing, interpreting, and discovering with the FinnGen data [2]. The FinnGen Analyst Handbook is available for FinnGen researchers at Members Area [3] on FinnGen web pages [2].

FinnGen project is one of the biggest Biobank research projects worldwide [2]. There are almost 900 researchers with FinnGen account at Finnish Biobanks, Universities and Hospitals and at the 13 partner pharmaceutical companies. One goal of the Analyst Handbook was to enhance research efficiency by increasing the exchange of knowledge between FinnGen researchers and by reducing the researchers time wasted looking for information, tools, and data.

This Master's Thesis work was to write documentation for FinnGen Analyst Handbook. The work included studying the FinnGen project, learning to use all FinnGen custom-made tools well, producing template codes for FinnGen researchers to apply in their work, and writing instructions in the Analyst Handbook. The author of this thesis contributed to the FinnGen project by writing sections to Handbook, testing FinnGen programs, answering FinnGen helpdesk questions, and conducting FinnGen admin tasks like security checking of files. This thesis gives a report about the Analyst Handbook and its writing process. In addition, one example of the entire workflow for GWAS using Analyst Handbook instructions, FinnGen custom-made tools, and R coding is provided.

## 2 BACKGROUND

### 2.1 FinnGen Analyst Handbook

One of the main ideas behind FinnGen Analyst Handbook was to provide a guide for FinnGen partner researchers on how to use the FinnGen secure environment Sandbox [4] and FinnGen custom made tools and general analysing tools within Sandbox. All analyses, performed by FinnGen partner researchers, using FinnGen data are conducted in the secured environment called FinnGen Sandbox [4]. One goal of Analyst Handbook is to enhance research efficiency by reducing the time wasted looking for data and analysis tools. FinnGen researchers come from various backgrounds including clinicians, biologists, statisticians, bioinformaticians, and data analysts. The goal of the Analyst Handbook is to serve all these users. A lot of information about genetics, statistics, registers, and methodology is also provided in the Handbook to fill in the gaps of what users may need to know about FinnGen project. The landing page of the FinnGen Analyst Handbook with the navigation bar is given in Figure 1.

The FinnGen analyst Handbook is jointly written by FinnGen staff in nine teams. These teams contain experts in bioinformatics, data analysts, program developers, clinicians, and administration. Each writer wrote sections regarding their own special fields. In addition, many of the topics in the Handbook was picked up from users' questions that were answered by FinnGen staff members on FinnGen Slack channel or at FinnGen Helpdesk. The FinnGen Documentation team organized the documentation collection, and Handbook structure, and wrote many of the sections. The Analyst Handbook is constantly updated with new topics arising from users' questions and coming along FinnGen data and tool development.



**FinnGen Analyst Handbook**

Search...

**Introduction**

Welcome to the FinnGen Analyst Handbook!

The following documentation contains both useful background information and clear instructions that are meant to help you get the most out of FinnGen data.

After a one year embargo, the FinnGen research project regularly publishes data releases containing summary statistics and Genome Wide Association Study results, which can be freely used. Additionally, researchers in organisations that have partnered with the FinnGen study can request access to either newest aggregate results or, when wishing to do their own analyses, access to the FinnGen Sandbox environment.

If you are using publicly available FinnGen data or ready results published by the FinnGen analysis team, you will find main concepts and data specifics useful. If you are a FinnGen partner organisation affiliate and are or will be working on your own analysis inside the FinnGen Sandbox, you might also be interested in reading further into the section **'Working in the Sandbox'**.

**How FinnGen data is processed**

The handbook contains these main topics:

1. [Main Concepts](#)
2. [FinnGen Data Specifics](#)
3. [Working in the Sandbox](#)
4. [Frequently asked questions](#)
5. [Release Notes](#)

**Figure 1.** The landing page of the FinnGen Analyst Handbook. The left panel gives the navigation table opened from Background Concepts to reveal the lower level titles. The right panel gives the landing page with the navigation table opened to show titles under FinnGen Data Specifics and Working in the Sandbox sections. Many of the lower-level titles are not shown. All titles, publication dates, and version numbers for 290 pages in the Analyst Handbook are given in Appendix Table 1.

## 2.2 FinnGen research project

FinnGen is a research project that brings together Finnish universities, Finnish hospitals, and hospital districts, Finnish institute for health and welfare (THL), biobanks, and international pharmaceutical companies [2]. Eleven Finnish biobanks, established by universities, hospital districts, and other research organizations, collect and provide the samples [1,2,5]. Nine of these are also FinnGen partners. FinnGen partner biobanks are Auria Biobank, Helsinki Biobank, Hematological Biobank (FHRB Biobank), Biobank of Eastern Finland, Central Finland Biobank, Northern Finland Biobank Borealis, Finnish Clinical Biobank Tampere, THL Biobank, and Blood Service Biobank [2]. Biobanks participating to the samples of the FinnGen, study but who are not official FinnGen partners are Arctic Biobank (University of Oulu) and Terveystalo Biobank Finland [2]. The sample collection is coordinated by the Helsinki Biobank and the University of Helsinki is the official data controller of the study. Finnish institute for health and welfare (THL) is responsible for handling and processing the register data.

The FinnGen study is funded by Business Finland and thirteen international pharmaceutical companies: Abbvie, AstraZeneca, Boehringer Ingelheim, Biogen, Bristol-Myers Squibb, Genentech, a member of the Roche Group, GlaxoSmithKline (GSK), Janssen, Maze Therapeutics, MSD (the tradename of Merck & Co., Inc, Kenilworth, NJ USA), Novartis, Pfizer, and Sanofi [1,2].

FinnGen contains three phases [2]. The first FinnGen phase from Aug 2017 to Aug 2020 included data collection and construction of infrastructure. The second phase from Aug 2020 to Aug 2023 continues data collection and infrastructure development and includes the first phase of analysis. At the time of writing this thesis in April 2022, FinnGen is in phase 2. Phase 3 is under planning. Phase 3 from Aug 2023 to Aug 2027 is planned to be the main analysis phase and also includes some new functional profiling data and analysis.

FinnGen has reached phase two (2020-2023) where 71 % of the data is collected and analysed. FinnGen releases data two times a year into a secure environment called FinnGen Sandbox [4]. The Sandbox provides a secure environment for researchers in universities and partner organizations to conduct analyses on the FinnGen data. The current data freeze DF9 contains 392,000 participants.

To provide researchers, clinicians, and statisticians secure access to the pseudonymized genome and digital health care data FinnGen has developed a secure environment called FinnGen Sandbox [4]. Sandbox is developed by a third-party contractor (currently Solita), which is a Finnish IT company [6]. Through Sandbox, FinnGen partner researchers may conduct analyses (e.g., Genome-Wide Association Study, GWAS) with pseudonymized

patient genetic and health care data. FinnGen Sandbox contains the FinnGen data and several analysing tools. Sandbox has custom tools as well as widely used analysing tools. Researchers will conduct all their analyses inside the secure environment of Sandbox [4]. Export of summary data and figures e.g., for publication is possible after a security check that guarantees the exported data doesn't contain individual level data [1]. As FinnGen is a research project all analyses conducted in FinnGen aim to be published. FinnGen researchers focus to find genetic associations for conditions, develop diagnostics, and personalized treatments for a wide variety of human diseases and conditions.

FinnGen also provides GWAS summary statistics for several endpoints ( $n = 4656$  in DF9) for FinnGen partner researchers. FinnGen endpoints are diseases and health-related conditions that are based on the health registry data and designed by FinnGen clinical, register, and analysing teams. Risteys is freely available tool for browsing FinnGen Endpoints [7]. Aggregate level summary data (also called "green data") is available for FinnGen researchers to discover from the ready-made GWAS and phenome-wide association study (PheWAS) results. After the embargo of 12 months, core analyses results are made publicly available through FinnGen web pages [8] thus making them freely available to the wide global scientific community.

## 2.3 Description of Analyst Handbook users

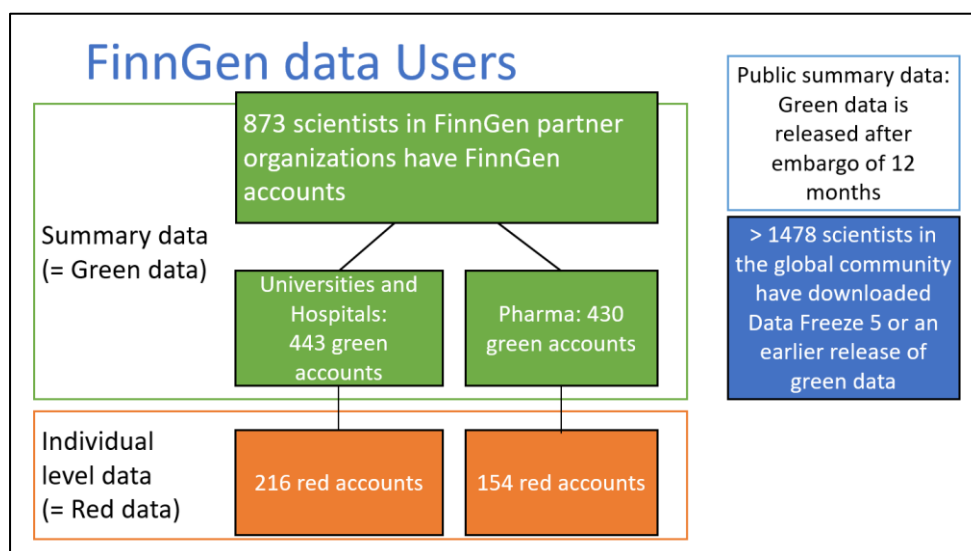
FinnGen data users come from various backgrounds. Within FinnGen data users there are clinicians, geneticists, biologists, bioinformaticians, biostatisticians, hard-core data analysts, computer scientists, and program developers. The aim of the Handbook is to serve FinnGen data users coming from these various backgrounds and provide the knowledge each user needs in their research. The Analyst Handbook [1] provides instructions on how to use the summary data (green data) and green data tools. For individual-level data (red data) users, the Analyst Handbook provides instructions on how to conduct analyses using FinnGen Sandbox and tools therein and provides many example codes and coding tips. Sandbox was initially designed for data scientists with good coding skills. Later FinnGen custom tools with a graphical user interface (GUI) were also added to the Sandbox. With FinnGen custom tools no coding skills are needed to conduct analyses but if users prefer to use their own coding and programs within Sandbox it is also possible.

Figure 2 gives the numbers of FinnGen data users in September 2021. In total 873 scientists had a FinnGen account. Green data includes summary data (non-identifiable data) like GWAS results for FinnGen endpoints (see FinnGen research project, p. 4). Out of green data users,

around half (443, 50.7 %) are at Finnish Universities or Hospitals, and around half (430, 49.3 %) at pharmaceutical partners.

Scientists who want to conduct analyses independently need access to individual-level data i.e. the red data. All red data users have also green data access. Around 58 % of FinnGen data users have green data access and 42 % have both red and green data access. Red data contains pseudonymized health care data and imputed and raw genotype data. Red data accesses are divided as 216 (58 %) and 154 (42 %) for Finnish Universities and Hospitals and for pharmaceutical partners, respectively (Figure 2).

FinnGen Analyst Handbook [1] serves both FinnGen Red and FinnGen Green data users. FinnGen Analyst Handbook is available in FinnGen Members Area [3] where users can register with their FinnGen account. Thus, at the time of writing this Master Thesis, the FinnGen Analyst Handbook is only for partners and does not serve users without a FinnGen account. However, there has been discussion about whether it is possible to publish parts of the Analyst Handbook and make it available for all users in the global scientific community.



**Figure 2.** Numbers of FinnGen data users in September 2021. On the left are numbers of scientists with FinnGen account at Finnish Universities and Hospitals or at partner pharmaceutical companies. On the right are users of public Summary data sets that is green data published after an embargo of 12 months. Handbook users are FinnGen researchers with FinnGen accounts.

## 3 MATERIALS AND METHODS

### 3.1 The growing project needed detailed documentation

In the early years of the FinnGen project documentation was sparse. The growing FinnGen project needed detailed documentation that is accessible to all FinnGen data users and FinnGen staff. One source of the truth would ease research by offering instructions and template codes and save time spent on searching the data and tools.

After Data Freeze 4 users started to ask for instructions on how to make analyses. Roadmap meetings were organized to plan users' guidance. Online FinnGen users' meetings were initiated with demos on how to use FinnGen tools. Users' meetings were recorded, and recordings were set first to FinnGen SharePoint for users to watch. Soon after, Members' Area web page was established. Members Area contains internal documents, meeting recordings, and tutorials for FinnGen data users. Later Users' Meetings recordings were also loaded to the Members' Area. However, the need for written instructions remained.

First idea of the FinnGen Analyst Handbook was planned by the leader of the FinnGen Trajectory team Mary Pat Reeve and Pinja Krook the service designer at Solita (see History of the FinnGen Analyst Handbook – from concept to implementation, p. 11). The first version of the list of content for the Handbook was planned. User interviews were conducted (see User interviews, p. 10). Users were asked what they would expect to find under the pullet points of the list of content. Based on the users' comments and suggestions the list of content was reviewed. This made the first backbone where the writing of the Analyst Handbook later started (see History of the FinnGen Analyst Handbook – from concept to implementation, p. 11). Other Handbooks were explored for inspiration. These Handbooks were Alicia Martin at Broad [9], Open Targets [10], United Kingdom Biobank (UKBB) [11], and Ensembl [12]. A documentation team of six people was created to coordinate the Handbook writing process (see Teams who wrote the FinnGen Analyst Handbook, p. 8).

## 3.2 Teams who wrote the FinnGen Analyst Handbook

All FinnGen teams participated in the writing of the FinnGen Analyst Handbook. The FinnGen Documentation Team, including the author, organized the collection of the text and other data. The documentation team found writers for each section within the team or asked specialists in other teams to write sections on their expertise. Documentation team members wrote many of the Handbook sections and conducted editing and proofreading. The writing process of the FinnGen Analyst Handbook was a joint effort of all FinnGen teams. At the time of writing this Master Thesis in April 2022 there are 31 writers in the FinnGen Analyst Handbook. The sections titles of the Analyst Handbook are given in Appendix Table 1.

The FinnGen teams are

FinnGen Admin Team (8 members)

FinnGen Clinical Team (6 members)

FinnGen Register Team (7 members)

FinnGen Sequencing Informatics Team (5 members)

FinnGen e-Science Team (7 members)

FinnGen Trajectory Team (5 members)

FinnGen Analysis Team (7 members)

Data Science - Genetic Epidemiology Lab (4 members)

Documentation Team (7 members)

## 3.3 Technology

Several options for the platform of the Handbook were considered: Google documents [13], DocuSaurus inside Sandbox [4], Members Area [3], GitBook [14], and GitHub [15]. Project management and communication tools Wrike [16] and Slack [17] were used to manage the Analyst Handbook writing process.

The first draft of the FinnGen Analyst Handbook was collected in Google documents [13]. This was planned as a temporary solution to enable to start of the collection of documentation even though the final platform was still under consideration. The list of content of the Handbook was set to Google documents to give a structure for the documentation. FinnGen staff members writing the Handbook sections included their sections to the Google documents. Meanwhile, the design of the permanent platform for the FinnGen Analyst Handbook continued.

DocuSaurus is software that was built inside FinnGen Sandbox by Solita [6] to hold the documentation. However, DocuSaurus was quickly discontinued due to large overhead for updating documentation, lack of real-time upgrades, and difficulties including images or linking videos to the documentation (see History of the FinnGen Analyst Handbook, p. 11). In addition, there are two kinds of users in FinnGen: red and green data users from which only the first one has access to FinnGen Sandbox. The Analyst Handbook contains a lot of information for all FinnGen data users, also for those with green data access. It was considered to split Handbook into two separate books: sections for green data users set in Members Area [3] and sections for red data users in DocuSaurus in Sandbox. After the first draft version of the Analyst Handbook started to get content, it became clear that there will be a lot of linking between pages throughout the Handbook. Splitting Handbook into two books was no longer an option.

Members Area is a web page for FinnGen partners [3]. It was not built for documentation in a book-like format and was not able to hold Analyst Handbook content. However, a page where the Analyst Handbook is published was later created in Members' Area (see Publishing of the FinnGen Analyst Handbook, p. 15).

Based on the previous experience with GitBook [14] it was selected as the platform for the FinnGen Analyst Handbook (see History of the FinnGen Analyst Handbook p. 11). Gitbook has options to create teams with admin, writer, and reader permissions which was beneficial for this kind of documentation having several writers. GitBook updates immediately and its usage was simpler than for the other options. The Analyst Handbook documentation was moved from Google documents to GitBook by the documentation team.

The FinnGen analyst Handbook has two versions. One is the production version of the Analyst Handbook that is available to FinnGen data users at FinnGen Members Area. The other one is a draft version of the Handbook containing all the content in the production version and new documentation that is under the preparation or proofreading and editing phase. Approximately once a month the Handbook is updated.

To update the Handbook GitBook was synchronized with GitHub [15] repository. GitBook content of the Handbook draft version was pushed to GitHub repository. Then the content from GitHub repository was pulled to GitBook space of the Handbook production version. The GitHub repository step made it possible to maintain two versions of the FinnGen Analyst Handbook. GitHub repository also provides backups and version control of the Analyst Handbook (see the version history of the Analyst Handbook, p. 13). Updating Handbook and managing GitBook spaces and GitHub repositories was on responsibility of the author.

Project management tool Wrike [16] was used to keep track of Analyst Handbook sections under preparation, assign tasks to writers, and follow-up sections completing. FinnGen community Slack [17] was used to inform users about coming updates and maintenance breaks. Several Slack questions and answers were included in the Analyst Handbook. Slack was also frequently used for communication within FinnGen documentation team. Coming updates and maintenance breaks of FinnGen Analyst Handbook were also announced to data users with e-mailing lists.

### 3.4 User interviews

Two interviews for Handbook users were arranged by FinnGen and Solita (Solita [6] is FinnGen/FIMM subcontractor company that develops FinnGen Sandbox environment) to improve users' experience of the Analyst Handbook.

The first interview was conducted in late 2020 after the table of contents was established in September 2020. Service designer Pinja Krook from Solita and Trajectory team leader Mary Pat Reeve from FinnGen conducted seven 30 min user interviews. At this point of the Handbook design, the content was outlined, and contents headings were drafted. Users were asked what kind of content they would expect to find under each page title and would that knowledge be beneficial for them. Users were asked what topics are not covered in the table of contents and which kind of information would benefit them most. Based on the user interview 97 topics of the Handbook pages were established and content designed further.

The second interview was arranged in March 2022 by FinnGen and Solita. Service designer Maiju Samberg from Solita interviewed nine users and compiled a report from the interviews. One of the FinnGen Documentation team members joined in every meeting. Notes from the interviews were discussed in Documentation team meetings. At the time of the second interview, Handbook contained 221 pages. The content for 97 pages designed after the first interview were ready and published by November 2021. In addition, 124 new pages were included in November's release.

In the second interview, users were asked how they use Handbook, what kind of content they are expecting to see in Handbook, and if this content is found in Handbook. The report of the interview showed that most users look to Handbook as their first source of help.

The most visited sections were Background Concepts, FinnGen Data Specifics, FinnGen Data Freezes and Releases, Detailed Longitudinal Data, Other registry data files in Sandbox, Working in the Sandbox, How to get started with Sandbox, Running analyses in Sandbox,



Custom GWAS command line (CLI) tool, How to run GWAS using REGENIE, How to run GWAS using SAIGE, FAQ, and Release Notes.

The most used search words were: REGENIE, SAIGE, ATLAS, GWAS, PHEWEB, DOCKER.

Taking together statistics of the most used pages and search words, most users are looking for instructions on how to conduct their own analysis. Users are also exploring the FinnGen data structure, the data that are available, and the data that are released. Knowing the FinnGen data structure is needed for conducting analyses but also for planning future research suggesting that users may also use Handbook to plan future studies on FinnGen data.

Based on the interviews and report by Maiju Samberg users find the Analyst Handbook very useful. Based on Maiju's suggestions and users' feedback the work to make Handbook even better continues.

### **3.5 History of the FinnGen Analyst Handbook - from concept to implementation**

As is often the case with a large start-up project, documentation for FinnGen was sparse in the early years. How much a user could accomplish in the secure Sandbox environment was often dependent on if they knew someone to ask for help. One goal of the Analyst Handbook was to democratize the ability to do analysis and reduce the time wasted looking for data and analysis tools. FinnGen users come from various backgrounds - some are clinicians still working in clinical settings, some are bioinformaticians, and some are hard-core data analysts. The goal of Analyst Handbook is to help all of these users fill in the gaps of what they needed to know about the FinnGen project.

During the first years of FinnGen, the main form of documentation was READMEs in the data directories and a deck of training slides with a companion video made by Solita [6]. At the time, there was no central place for FinnGen users to find the training video easily - many had trouble accessing FinnGen SharePoint and the FinnGen Members Area [3] did not yet exist. The video, a gestalt unit without breaks, was also very difficult to update as tools in the Sandbox rapidly evolved with each release.

Several other FinnGen updates helped move the documentation in a better direction. One was opening a Slack community to all FinnGen users. The Slack channels provided a way to get answers without knowing who to contact and record what questions people asked most frequently. (However, it also meant that FinnGen code developers were frequently interrupted to answer the same questions multiple times.) In 2019, FinnGen upgraded the publicly available information about data releases to include details of the analysis methods and

statistics on the releases [1]. All the information was gathered and organized it within the GitBook [14] platform to make FinnGen Documentation pages [18]. We also introduced the project management system, Wrike [16], an important step later in sourcing documentation efforts from all FinnGen team members. Pinja also invented the idea of “Individual Jones”, a take-off of Indiana Jones, to bring a storyline to the data security training videos that are now required viewing by all FinnGen partners. These security videos helped raise the baseline awareness of the FinnGen project for all Sandbox users.

Mary Pat Reeve, FinnGen Trajectory team leader, found in early 2020 an interesting whitepaper from GitLab about their “Handbook-first documentation” philosophy [19], that helped us formulate the goals of our documentation effort. GitLab describes Handbook-first documentation: “A handbook-first organization is home to team members who benefit from having a single source of truth to lean on. This type of organization is able to operate with almost supernatural efficiency. An organization that does not put a concerted effort into structured documentation has no choice but to watch its team members ask and re-ask for the same bits of data in perpetuity, creating a torturous loop of interruptions, meetings, and suboptimal knowledge transfers.” This whitepaper outlines procedures to make the “Handbook first” approach work, many of which were implemented with the FinnGen Analyst Handbook. Handbook first idea helped to formulate a clear set of goals, namely, a centralized source of all documentation, to empower all internal FinnGen teams to contribute via the handbook editor team, and to cover the building blocks as well as day-to-day procedures so that everyone would be able to use the data to its fullest potential.

Once the goals for the documentation were clear, the next step was to find a platform to house the documentation. In September 2020, Pinja, Mary Pat Reeve, and Mervi Aavikko, FinnGen Project Manager, began to look for a system to hold more detailed information on using the Sandbox environment. Security measures impose limits on copying and pasting to and from the Sandbox environment, so initial idea was to look at systems that could mirror the documentation into the Sandbox where users could easily paste code and path names. DocuSaurus system within Sandbox was implemented, but it was quickly discontinued after rollout due to large overhead for updating documentation (ten steps and three specialized software tools required), lack of real-time upgrades (only mirrored once daily), as well as difficulties including images or linking videos to the documentation. Since the FinnGen team already had a positive experience and in-house expertise with GitBook [14], it was decided to move forward with GitBook. An additional advantage of GitBook is the lower cost of backend maintenance than DocuSaurus.

In parallel with documentation tool testing, Mary Pat and Mervi outlined all the sections they would like to have in the documentation. Seven 30-minute interviews with users of different

backgrounds were arranged to ask what users would expect to find in each outlined category and adjusted the phrasing and sections accordingly (see Users interviews, p 11). Users were also encouraged to suggest any sections that might be missing. Mary Pat also worked Solita team on building a visual map of the Sandbox environment to be placed on the backdrop of the Sandbox as a reference for users.

In spring 2021, Susanna Lemmelä was appointed project owner for the Handbook editor team, and it was then christened “The FinnGen Analyst Handbook”. Susanna managed the project-management side of gathering various existing files of documentation, tasking users to create new documentation and Marianna Niemi managed the harmonization of data into GitBook [14] and the technical aspects of each release. The Handbook went live in June 2021.

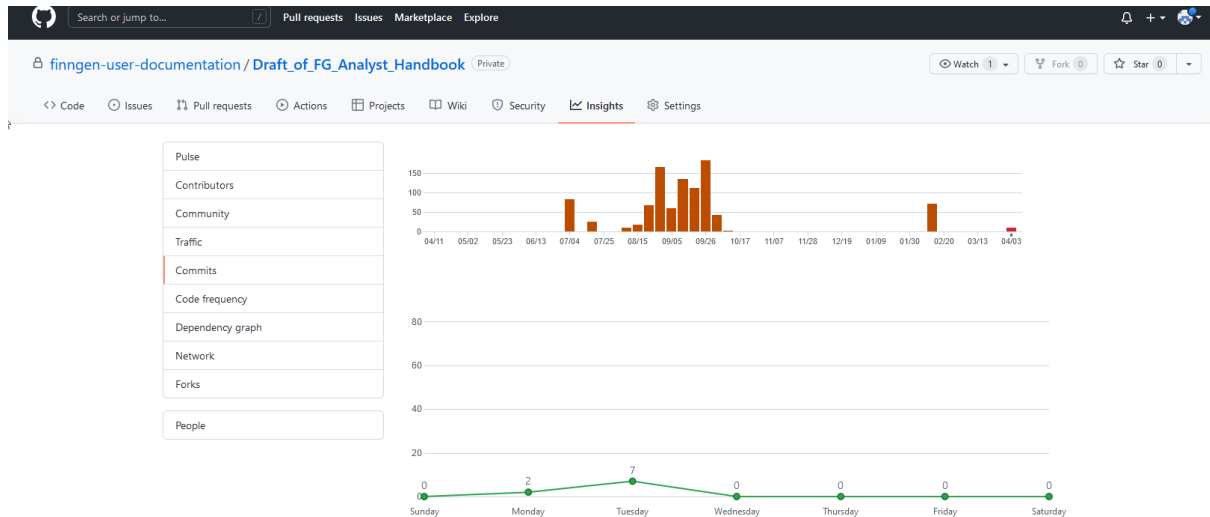
### 3.6 The version history of the Analyst Handbook

The FinnGen Analyst Handbook was first published online on the 15<sup>th</sup> of June 2021. The Analyst Handbook is an electronic guidebook that is updated regularly. By the time writing this thesis in April 2022, there have been seven updates to the Handbook and in total 290 pages (Table 1). Version control with Git including synchronizing between GitBook and GitHub, merging, managing GitHub branches, backups, and version history has been the author’s responsibility (see Technology, p. 8).

**Table 1.** Summary table of the FinnGen Analyst Handbook version history. The Version number, Date, Number of pages, and a cumulative number of pages of the Analyst Handbook updates are given. Detailed version History of the Analyst Handbook is given in Supplementary Table S1.

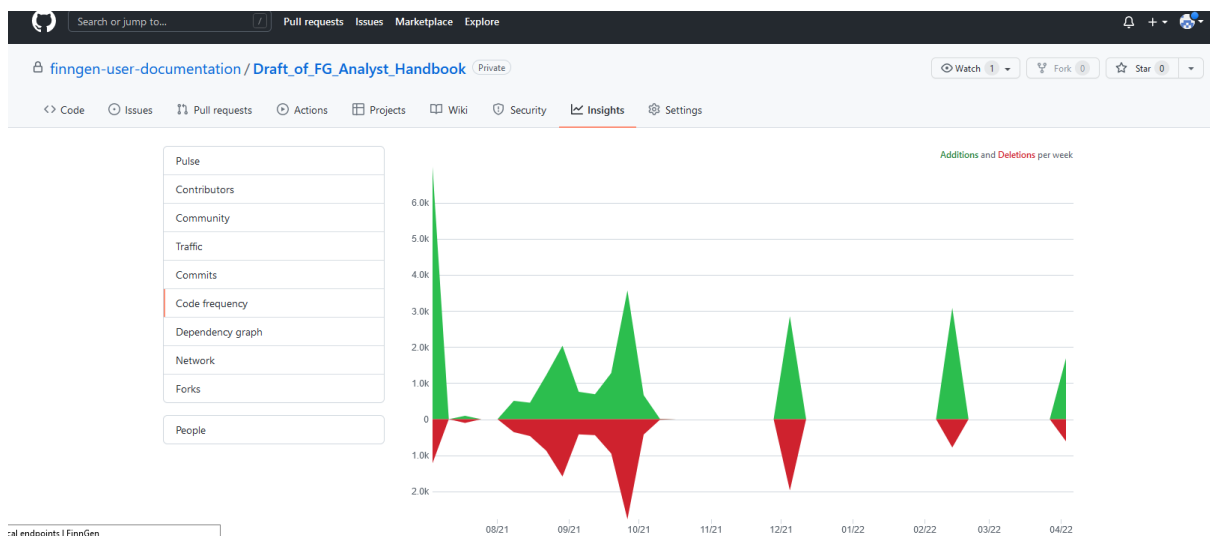
Version	Date	N pages	N pages cumulative
1	15 <sup>th</sup> June 2021	14	14
2	9 <sup>th</sup> July 2021	174	188
3	1 <sup>st</sup> October 2021	29	217
4	1 <sup>st</sup> November 2021	4	221
5	10 <sup>th</sup> December 2021	11	232
6	14 <sup>th</sup> February 2022	46	278
7	4 <sup>th</sup> April 2022	12	290

Commits to the GitHub [15] repository of the draft version of the Analyst Handbook shows the development of the Handbook (Figures 3 and 4). The first commit was in June 2021 when the first version of the Analyst Handbook was established. In the autumn 2021 Handbook was under heavy development and hundreds of sections were added. The following updates were done with fewer commits to the repository.



**Figure 3.** Commits to the GitHub repository of the draft version of the Analyst Handbook shows the development of the Handbook. The first commit was in June 2021. In the autumn 2021 Handbook was under heavy development. The following updates were done with fewer commits to the repository.

Code frequency in June 2021 reflects the first build-up of the Handbook showing more code adding (~8.0k in green) compared to code deletion (~1.0k in red, Figure 4). In September and October Analyst Handbook was under heavy development during which GitBook space and GitHub repository were constantly synchronized (Figure 4). Existing chapters were updated and ~1.0k of code for new chapters was added. A similar ~1.0k amount of new code was added in December 2021 and April 2022 updates (Figure 4). February 2022 update was a large update of ~2.0k of new code corresponding to 46 new sections (Figure 4, Table 1).

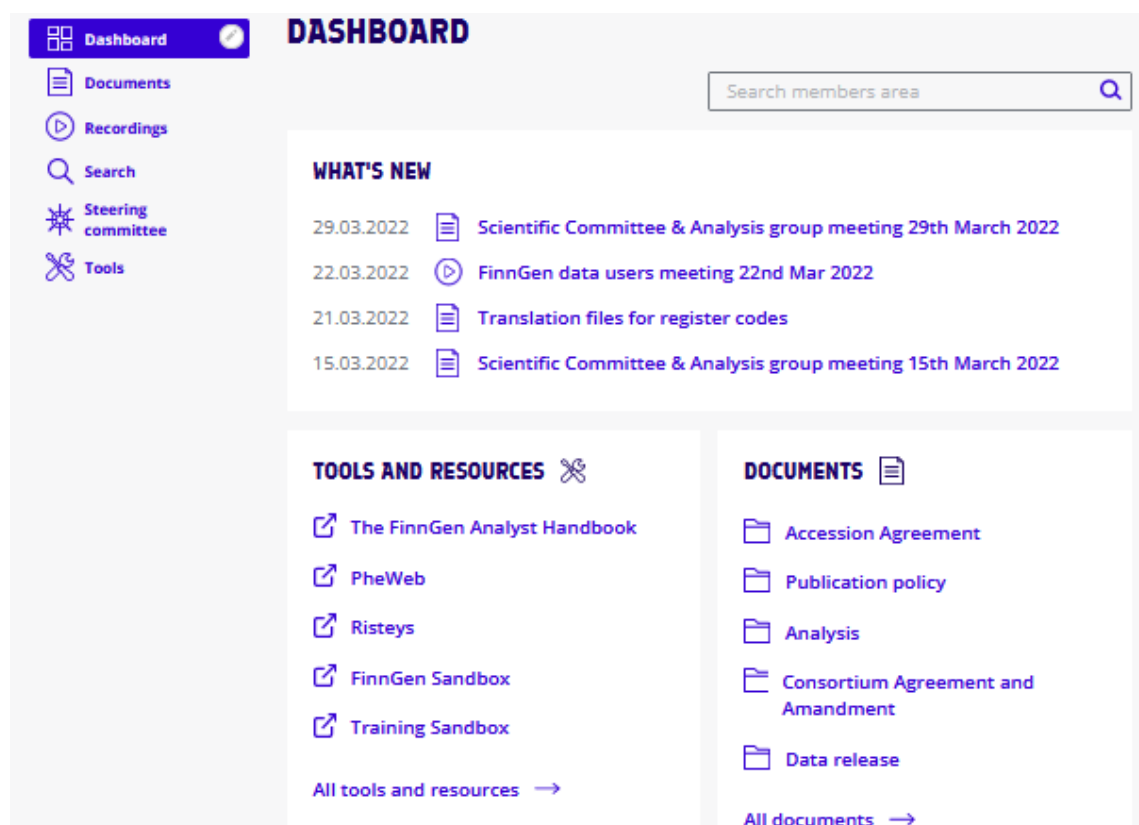


**Figure 4.** Code frequency of the GitHub repository of the draft version of the Analyst Handbook shows the size of code additions and deletions per week. Additions of code are indicated in green, and deletions of code are in red. Code deletion arose from page updates when new code replaces the old code. The amount of new code for new chapters can be roughly calculated from code added (green) minus code removed (red).

### 3.7 Publishing of the FinnGen Analyst Handbook

FinnGen Analyst Handbook is published as an electronic GitBook [14] document available for FinnGen researchers in FinnGen Members Area internet pages [3]. Members Area contains internal documents, meeting recordings, and tutorials for FinnGen data users. Accessing Members area needs a FinnGen account.

Handbook updates and new Handbook sections are announced in FinnGen Users' Meetings. Users' Meetings are organized usually twice a month. On the agenda are current topics, updates, announcements of new tools, and tutorials for using those tools. The Users' Meetings are recorded. After the meeting, the records are available for FinnGen researchers on FinnGen SharePoint pages. Recordings and meeting pdf files are downloaded also on Members Area [3] (Figure 5).



**Figure 5.** A screenshot from the FinnGen Members Area front page. Link to the FinnGen Analyst Handbook is available under the TOOLS AND RESOURCES section in FinnGen Members Area. Updating content to the Members Area along with Handbook was on responsibility of the author of this thesis.

## 4 RESULTS

### 4.1 Statistics of the FinnGen Analyst Handbook usage

Since the FinnGen Analyst Handbook was first published online on the 15<sup>th</sup> of June 2021 the visits to the page have increased constantly. During the first month after the Handbook publishing there was 1563 visits on the pages (June 2021, Figure 6). In March 2022 there was more than 4317 visits on the FinnGen Analyst Handbook pages. According to the users interviews the Analyst Handbook is the first source of help for FinnGen researchers (see User interviews, p. 10).



**Figure 6.** shows a screen capture from the number of Monthly visits on FinnGen Analyst Handbook pages taken on 11<sup>th</sup> of April 2022. The visit survey histogram is a build-in feature of the GitBook. The FinnGen Analyst Handbook was first published online on the 15<sup>th</sup> of June 2021 after the visits have increased. In March 2022 there was 4317 visits. The number of visits in April 2022 reflects only the first 11 days of April when the screenshot was taken.

## 4.2 Example on how to conduct a custom GWAS in FinnGen Sandbox using instructions of the Analyst Handbook

### 4.2.1 Example workflow and cohorts

#### Workflow

Here an example of the whole workflow for genome wide association study (GWAS) for one human condition is provided. The analyses are conducted using instructions given in the FinnGen Analyst Handbook [1]. The selection of these examples suits well to this Master Thesis work as also the Handbook sections needed for these analyses are written by the author of this thesis.

The workflow includes following steps:

1. Building cases and controls cohorts,
2. GWAS on the cohorts, and
3. GWAS results viewing with FinnGen PheWeb tool.

The cases and controls cohorts building and GWAS are conducted using two approaches:

- A. using FinnGen custom made tools: Atlas, and custom GWAS tools, working from drop-down menus and needing no coding skills, and
- B. using coding and FinnGen command-line tools: coding in R language and conducting custom GWAS from the command line

#### Data

The data from where the cohorts are built is **Detailed Longitudinal Data** of FinnGen Data Freeze 7. The Atlas provides a graphical user interface (GUI) tool to conduct the searches on Detailed Longitudinal Data (approach A above). In approach B, the detailed longitudinal data will be downloaded to RStudio where the coding takes place.

Detailed longitudinal data is pre-processed data by FinnGen Registry Team. Most of FinnGen data users start their analyses from Detailed longitudinal data either with their own coding (R, Python, Jupyter, Bash) or with FinnGen Sandbox custom-made tools (Atlas, custom GWAS). Detailed longitudinal data combines the Hospital Discharge Register, Finnish Cancer Register, Cause of Death Register, Drug Purchases Register, Drug Reimbursement Register, and Primary Care Register in longitudinal format. In the longitudinal format, medical records are combined into one table. Each person has as many rows in the longitudinal table as there are visits to Hospitals, Primary care, medicine purchases, or other events in the Medical Registers

for that person. Usage of Detailed Longitudinal Data helps researchers significantly as the prior data cleaning and checking steps are already conducted by FinnGen register team.

### **Cases and Controls cohorts**

The example for my thesis will be DrugWas (= GWAS for drug users) for antihistamines. The cases cohort is defined by persons using antihistamine medicines. ATC codes for antihistamines are all ATC codes starting with R06A excluding Cinnarizine R06AE02 and Levocetirizine R06AE08. Cinnarizine and Levocetirizine are no longer used as antihistamines. The cohorts are defined in collaboration with clinicians from Tampere University Hospital and the Pulmonary research team having participating clinicians in several Finnish University Hospitals.

Criteria for cases cohort are

- ATC codes starting with R06A for antihistamines excluding Cinnarizine R06AE02 and Levocetirizine R06AE08
- Restriction of Kela purchases records on or after 1<sup>st</sup> of January 1995

Criteria for controls cohort are

- No use of antihistamines ever
- Having medical records on or after 1<sup>st</sup> of January 1995

## 4.2.2 Example on how to create GWAS analysis with FinnGen custom tools

### **Building cases and control cohorts with Atlas tool**

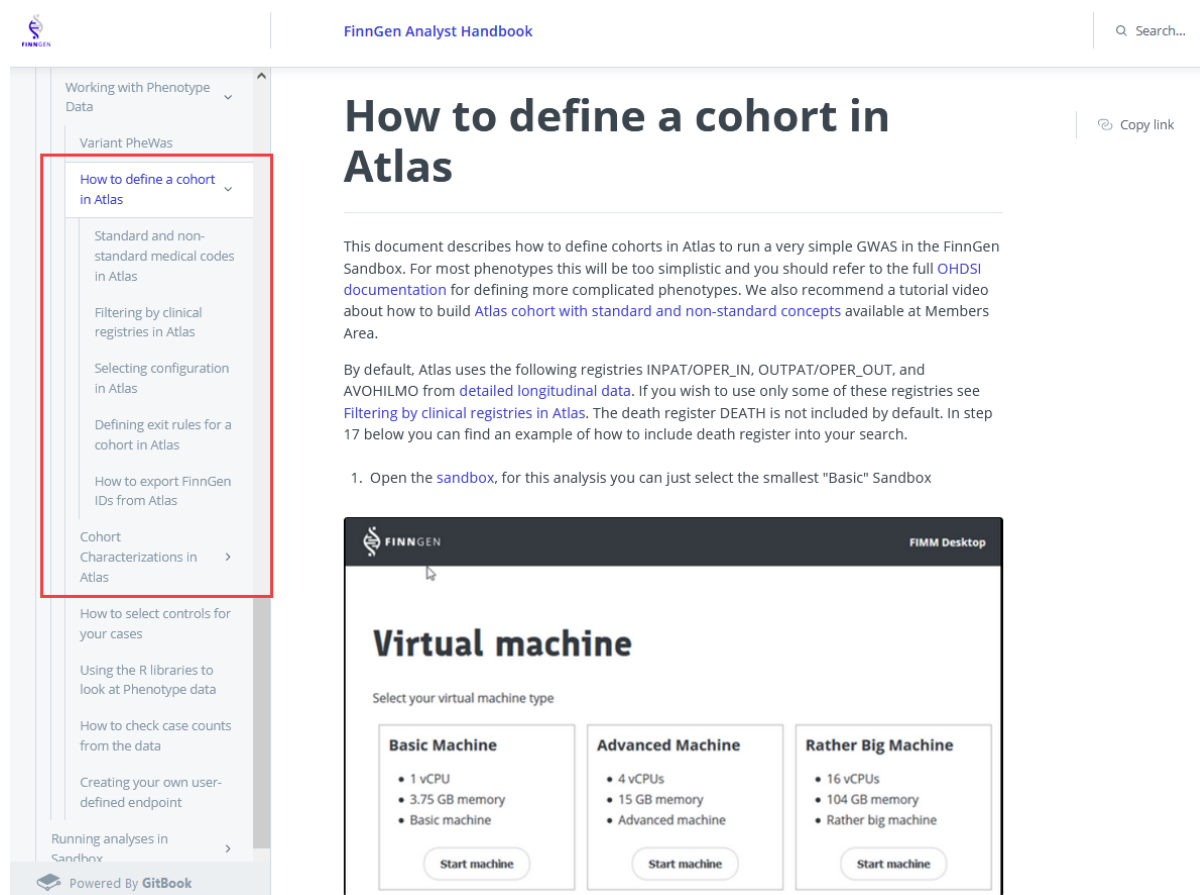
In this example cases and control cohorts are built with FinnGen tools Atlas and custom GWAS GUI tool. These tools have graphical user interface (GUI). Therefore, the usage of these tools needs no coding skills from the user.

The instructions for how to build a cohort in Atlas and how to run custom GWAS using custom GWAS tool are described in detail in the FinnGen Analyst Handbook [1] (Figure 7). Figure 1 shows the first Atlas section in FinnGen Analyst Handbook with Atlas sections highlighted. Atlas sections in the FinnGen Analyst Handbook gives detailed instructions on how to make cases and control cohorts, how to use standard and non-standard code sets, how to set inclusion and exclusion criteria, and how to visualize the cohorts with Atlas cohort characterization tool (Figure 1).

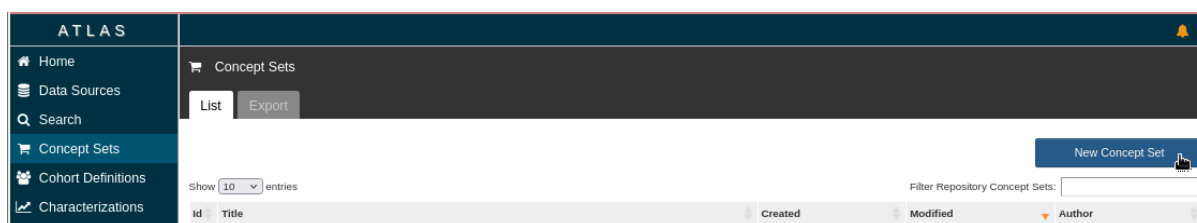
To build Atlas cohorts based on the antihistamine usage first step was to define a concept set for antihistamines (Figure 8). Concepts were selected using the Atlas Search option with the search word "R06A" which is the start of ATC codes for antihistamines (Figure 9). Resulting



90 codes were ordered by record counts (RC) and descendant record counts (DRC). All codes where RC or DRC were more than 0 were selected by clicking the shopping cart icon (Figure 9). In total 30 antihistamine medicines with any RC or DRC counts were included in the antihistamine concept set (Figure 8).



**Figure 7.** Screenshot from the FinnGen Analyst Handbook shows beginning of the section *How to Define a cohort in Atlas*. The other Atlas sections providing detailed instructions on how to create a cohort, use a different type of codes, filtering data, set configuration, set cohort exit rules, data exporting, and cohort characterizations are highlighted in red in the content panel on the right.



**Figure 8.** A concept set is created in Concept Sets page in Atlas by clicking “New Concept Set” button and giving a name for new concept set. Next concepts will be selected in concept sets using Atlas Search option (see Figure 9).

**FINN GEN** FIMM Desktop M

**ATLAS**

← Antihistamine\_R06A\_alwithRCorDRC[MN]

Search Import

R06A

Column visibility Copy CSV Show 45 entries Filter: Advanced Options

Showing 1 to 45 of 90 entries

	ID	Code	Name	Class	RC	DRC	Domain	Vocabulary
▼ Vocabulary								
ATC (30)	21003497	R06AE07	cetirizine; oral	ATC 5th	488,754	977,508	Drug	ATC
▼ Class								
ATC 5th (82)	21003526	R06AX27	desloratadine; oral	ATC 5th	216,399	432,798	Drug	ATC
ATC 4th (7)	21003521	R06AX22	etastine; oral	ATC 5th	150,440	300,880	Drug	ATC
ATC 3rd (1)	21003498	R06AE09	levocetirizine; oral	ATC 5th	128,935	257,870	Drug	ATC
▼ Domain								
Drug (90)	21003514	R06AX13	loratadine; oral	ATC 5th	114,038	229,276	Drug	ATC
▼ Standard Concept								
Classification (50)	21003525	R06AX26	fexofenadine; oral	ATC 5th	43,443	86,886	Drug	ATC
▼ Invalid Reason								
Valid (50)	21003513	R06AX12	terfenadine; oral	ATC 5th	9,428	18,856	Drug	ATC
▼ Has Records								
false (7)	21003512	R06AX11	astemizole; oral	ATC 5th	9,288	18,576	Drug	ATC
true (15)	21003500	R06AE53	cyclizine, combinations; systemic	ATC 5th	4,363	8,726	Drug	ATC
▼ Has Descendant Records								
false (27)	21003524	R06AX25	mizolastine; oral	ATC 5th	2,262	45,863	Drug	ATC
true (33)	21003493	R06AE03	cyclizine; systemic, rectal	ATC 5th	920	1,840	Drug	ATC
	21003495	R06AE05	medocline; oral, rectal	ATC 5th	262	2,786	Drug	ATC
	21003461	R06AB03	dimetindene; oral	ATC 5th	71	142	Drug	ATC
	21003460	R06AB02	deschlorpheniramine; systemic	ATC 5th	33	66	Drug	ATC
	21003454	R06AA52	diphenhydramine, combinations; systemic	ATC 5th	10	20	Drug	ATC
	21003468	R06AB54	chlorpheniramine, combinations; systemic	ATC 5th	0	5,202,540	Drug	ATC
	21003457	R06AA57	diphenhydramine, combinations; systemic	ATC 5th	0	4,361,737	Drug	ATC
	21003489	R06AD52	promethazine, combinations; systemic	ATC 5th	0	2,448,831	Drug	ATC
	40254475	R06AA39	doxylamine, combinations; systemic	ATC 5th	0	2,440,427	Drug	ATC
	21003445	R06A	ANTHISTAMINES FOR SYSTEMIC USE	ATC 5th	0	2,370,954	Drug	ATC
	21003466	R06AB51	brompheniramine, combinations; systemic	ATC 3rd	0	2,350,314	Drug	ATC
	21003467	R06AB52	dexchlorpheniramine, combinations; systemic	ATC 5th	0	2,263,844	Drug	ATC
	21003469	R06AB56	dexbrompheniramine, combinations; systemic	ATC 5th	0	2,202,095	Drug	ATC
	21003499	R06AE31	lucizine, combinations; systemic	ATC 5th	0	2,197,188	Drug	ATC
	21003455	R06AA54	clemastine, combinations; systemic	ATC 5th	0	2,053,668	Drug	ATC
	21003491	R06AE	Piperazine derivatives	ATC 5th	0	1,763,506	Drug	ATC
	21003503	R06AX	Other antihistamines for systemic use	ATC 4th	0	1,240,568	Drug	ATC
	21003456	R06AA56	chlorphenoxamine, combinations; systemic	ATC 4th	0	1,109,660	Drug	ATC
	21003501	R06AE35	medocline, combinations; systemic	ATC 5th	0	679,540	Drug	ATC
	21003517	R06AX17	ketotifen; oral	ATC 5th	0	14,642	Drug	ATC
	21003519	R06AX19	azelastine; oral	ATC 5th	0	11,030	Drug	ATC
	21003458	R06AB	Substituted alkylamines	ATC 5th	0	792	Drug	ATC
	21003446	R06AA	Aminoalkyl ethers	ATC 4th	0	86	Drug	ATC
	21003502	R06AK	Combinations of antihistamines	ATC 4th	0	0	Drug	ATC
	21003479	R06AD	Phenothiazine derivatives	ATC 4th	0	0	Drug	ATC
	21003470	R06AC	Substituted ethylene diamines	ATC 4th	0	0	Drug	ATC
	21003480	R06AD01	alimemazine; systemic	ATC 4th	0	0	Drug	ATC

Previous 1 2 Next

Apache 2.0  
open source software  
provided by  
**OHDSI**  
Join the Journey

**Figure 9.** Concepts were selected using the Atlas Search option with the search word “R06A”. Resulting 90 codes were ordered by record counts (RC) and descendant record counts (DRC). All codes where record counts were more than 0 were selected resulting 30 antihistamine medicines in the antihistamine concept set by clicking the shopping cart icon.

After concept sets were defined cases and controls cohorts were built using Atlas Cohort Definitions. For the cases cohort, a drug exposure to Antihistamine was defined by importing Antihistamine Concept Sets created above to Drug Source Concept (Figure 10). Records of Kela Medication purchases were restricted on or after 1<sup>st</sup> of January 1995 (Figure 10). Finally, the search was conducted on Detailed Longitudinal Data with given definitions on FinnGen Data Freeze 7 (Figure 11). The cohort was built on the Generation page in the Atlas tool by clicking Generate button (Figure 11). For FinnGen Data Freeze 7 in total 141737 persons with at least one purchase of antihistamines since 1995 were found (Figure 11).

**ATLAS**

Home | Data Sources | Search | Concept Sets | **Cohort Definitions** | Characterizations | Cohort Pathways | Incidence Rates | Profiles | Estimation | Prediction | Jobs | Configuration | Feedback

**Cohort #634**

antihistamin\_users[MN]

Definition | Concept Sets | Generation | Reporting | Export | Messages 3

enter a cohort definition description here

**Cohort Entry Events**

Events having any of the following criteria:

+ Add Initial Event

a drug exposure of Any Drug + Add attribute...

✗ occurrence start is: On or After 1995-01-01

✗ Drug Source Concept is: Antihistamine\_R06A\_allwithRC...

+ Delete Criteria

with continuous observation of at least 0 days before and 0 days after event index date

Limit initial events to: all events per person.

Restrict initial events

**Figure 10.** Cohort Definition settings for cases cohort of antihistamine users. Records of Kela Medication purchases were restricted on or after 1<sup>st</sup> of January 1995. Drug Source Concept is set to Antihistamine Concept Set as defined in Figure 9.

**ATLAS**

Home | Data Sources | Search | Concept Sets | **Cohort Definitions** | Characterizations | Cohort Pathways | Incidence Rates | Profiles | Estimation | Prediction | Jobs | Configuration | Feedback

**Cohort #634**

antihistamin\_users[MN]

Definition | Concept Sets | **Generation** | Reporting | Export | Messages 3

**Available CDM Sources**

	Source Name	Generation Status	People	Records	Generated	Generation Duration
<a href="#">Generate</a>	FinnGen CDM R6	n/a	n/a	n/a	n/a	n/a
<a href="#">Generate</a>	FinnGen CDM R7	COMPLETE	141,737	141,737	04/12/2022 9:57 AM	00:01:37 <a href="#">View Reports</a>
<a href="#">Generate</a>	FinnGen CDM R8	n/a	n/a	n/a	n/a	n/a
<a href="#">Generate</a>	FinnGen CDM R9	n/a	n/a	n/a	n/a	n/a

**Figure 11.** Conducting the search on Detailed Longitudinal Data with given definitions on FinnGen Data Freeze 7 (FinnGen CDM R7). The cohort was built by clicking Generate button on the Generation page in the Atlas tool. For FinnGen Data Freeze 7 in total 141737 persons with at least one purchase of antihistamines since 1995 were found.

The control cohort was built by first including Any Visit of any reason since 1995 in the Cohort Entry Event box on the Atlas Cohort Definitions page (Figure 12). Then antihistamine users were filtered out from the cohort by defining exactly zero occurrences of any drugs in

Antihistamine Concept Sets defined with Drug Source Concept option (Figure 12). As different medical registers may contain similar codes meaning different things the search was limited to the Kela Purchases register by selecting Visit occurrence to FinnGen Kela purchases (Figure 12). Finally, the search was conducted on Detailed Longitudinal Data on the Generation page in the Atlas tool by clicking Generate button (Top menu at Figure 12). For FinnGen Data Freeze 7 in total 178152 persons having medical records since 1995 but with zero purchases of antihistamines were found.

**FINNGEN** FIMM Desktop M

**ATLAS**

Home Data Sources Search Concept Sets Cohort Definitions Characterizations Cohort Pathways Incidence Rates Profiles Estimation Prediction Jobs Configuration Feedback

**Cohort #712**

NoUseOf\_antihistamins[MN]

Definition Concept Sets Generation Reporting Export Messages

enter a cohort definition description here

**Cohort Entry Events**

Events having any of the following criteria:

+ Add Initial Event

a visit occurrence of Any Visit + Add attribute... Delete Criteria

✗ occurrence start is: On or After 1995-01-01

with continuous observation of at least 0 days before and 0 days after event index date

Limit initial events to: earliest event per person.

Restrict initial events

**Inclusion Criteria**

New inclusion criteria

no\_antihistamins Copy Delete

1. no\_antihistamins  
Persons who haven't use antihistamines at all

Persons who haven't use antihistamines at all

having all of the following criteria: + Add criteria to group... Delete Criteria

with exactly 0 (using all) occurrences of:

a drug exposure of Any Drug + Add attribute...

✗ Drug Source Concept is: antihistamines[MN]

✗ with a Visit occurrence of: FinnGen Kela purchase Add Import

where event starts between All days Before and All days After index start date add additional constraint

☐ restrict to the same visit occurrence

☐ allow events from outside observation period

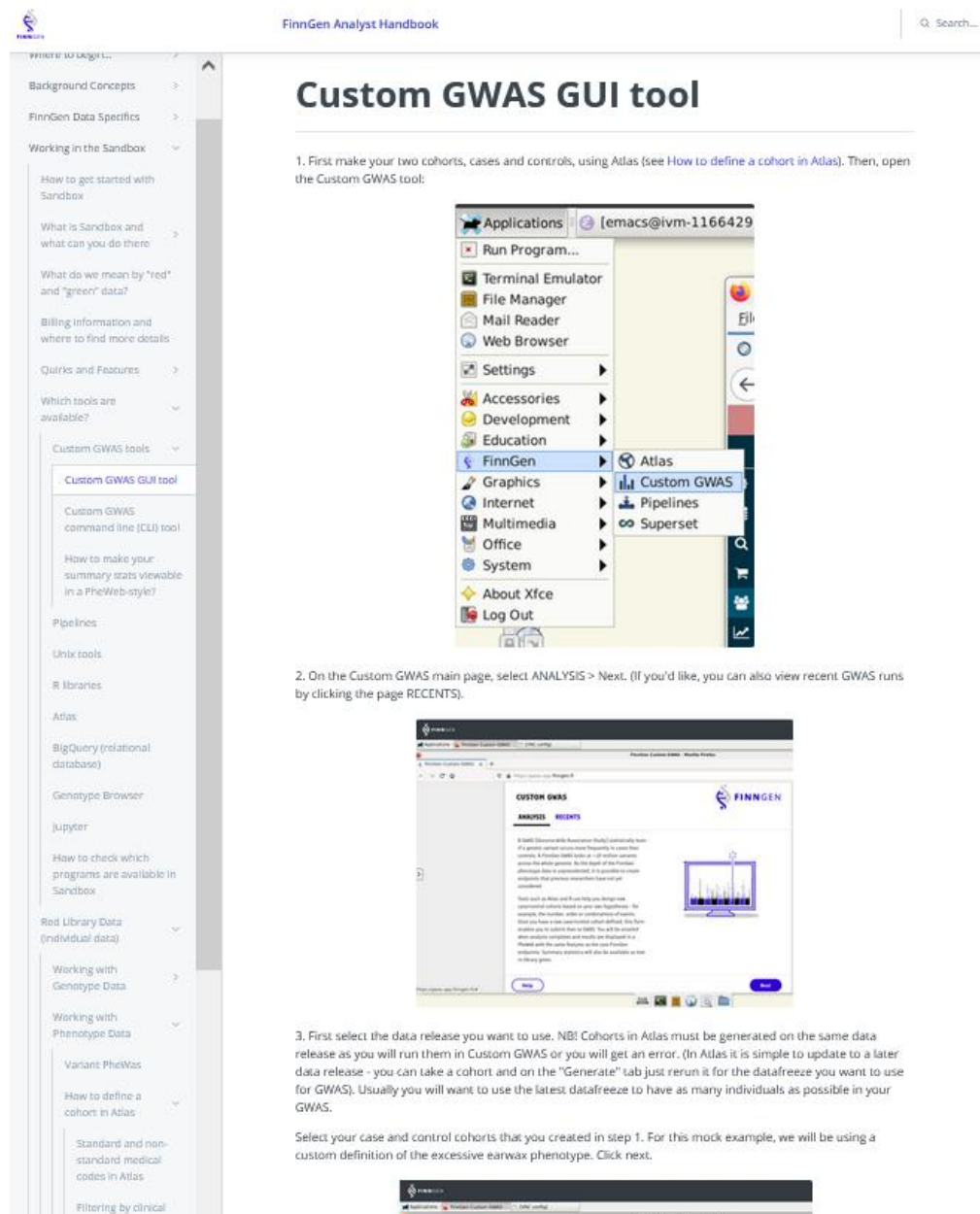
Limit qualifying events to: all events per person.

Apache 2.0  
open source software  
provided by

**Figure 12.** Cohort Definition settings for control cohort of persons never purchased antihistamines according to Kela Purchases records. First any visits were restricted to on or After the 1<sup>st</sup> of January 1995. Antihistamine users were then filtered out by defining exactly zero occurrences of any drugs in Antihistamine Concept Sets with code source limited to FinnGen Kela purchases registry. Drug Source Concept is set to Antihistamine Concept Set as defined in Figures 8 and 9.

## Conducting GWAS with custom GWAS GUI tool


Genome-wide association analyses were run on cases and control cohorts with FinnGen custom GWAS tool using FinnGen Handbook instructions (Figure 13). Custom GWAS GUI tool launched from the Application menu in FinnGen Sandbox is very easy to use for all users needing no coding skills at all (Figure 13).



**Figure 13.** A screen capture from “Custom GWAS GUI tool” section in FinnGen Analyst Handbook. The section in the Handbook was written by the author of this thesis.

FinnGen data release was set to Data Freeze 7 (Figure 14). Cases and Controls cohorts created in Atlas were selected from the drop-down menus of the Custom GWAS tool (Figure 14).

## CASES AND CONTROLS



### 1. Select a FinnGen data release

Different databases contain separate FinnGen data releases.

FinnGen CDM 7 (321,464 individuals)

### 2. Select cohorts

If you can't find suitable cohort, create a new one in Atlas.

#### Case cohort

antihistamin\_users[MN]

#### Control cohort

NoUseOf\_antihistamines[MN]

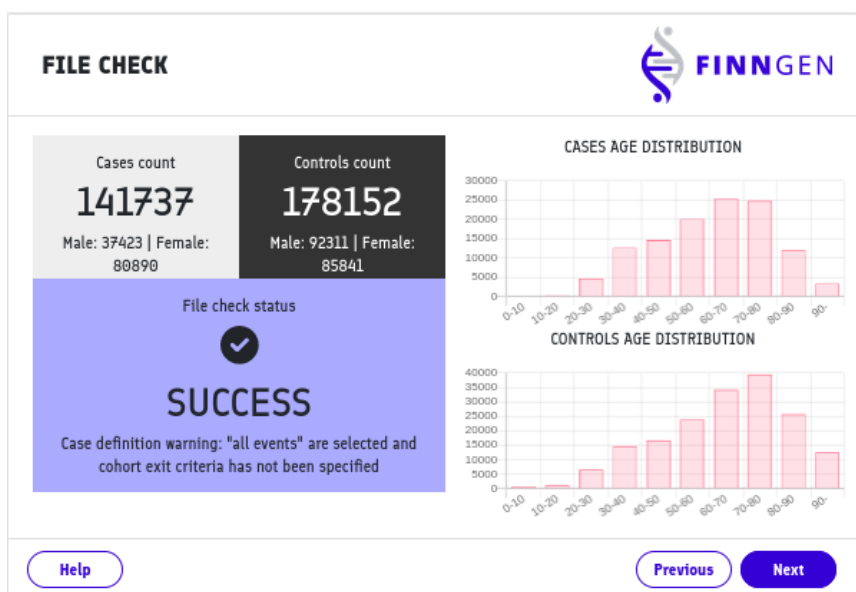
Standard covariates will be used in the analysis:  
Age, sex, 10 principal components, genotyping batch.

Help

Previous

Next

**Figure 14.** FinnGen Custom GWAS tool settings for antihistamine GWAS. Cases and Control cohorts built with Atlas were selected as Cases and Control cohorts in Custom GWAS tool setting page. Desired data release for GWAS run was set to Data Freeze 7.



**Figure 15.** FinnGen Custom GWAS tool gives summary statistics for cases and controls before launching GWAS. Cases and controls count, male and female counts in both cohorts and age distribution of the cases and controls cohorts.

## 4.2.3 Example how to build patient cohorts in R and conduct GWAS from the command line

### **Building cohorts with R – providing R scripts for cohorts building and visualization**

## R Notebook

Code ▾

Load libraries

Hide

```
library(tidyverse)
library(data.table)
library(plyr)
library(R.utils)
library(ggplot2)
```

Load Detailed Longitudinal Data

Hide

```
longitudinal_data_DF7 = fread("/finngen/library-red/finngen_R7/phenotype_2.0/data/finngen_R7_detailed_longitudinal.t
xt.gz", data.table = FALSE)
```

```
|-----|
|=====|
|-----|
|=====|
```

Define CASES GROUP:

Antihistamine users

- ATC codes starting with 'R06A'
- having records of buying R06A antihistamines after 1st Jan 1995 (modern antihistamines are in use)
- Kela drug purchase registry (PURCH) and Kela drug reimbursement registry (REIMB)

Hide

```
antihistUsers <- longitudinal_data_DF7 %>% filter(
  SOURCE == "PURCH" & (str_detect(CODE1, "^R06A") | str_detect(CODE2, "^R06A"))
)

# Separate years, days and months to columns and make them numeric
antihistUsers = separate(data = antihistUsers, col = APPROX_EVENT_DAY, into = c("EVENT_YEAR", "EVENT_MONTH", "EVENT_D
AY"), sep = "-", remove = FALSE)
antihistUsers$EVENT_YEAR = as.numeric(antihistUsers$EVENT_YEAR)
antihistUsers$EVENT_MONTH = as.numeric(antihistUsers$EVENT_MONTH)
antihistUsers$EVENT_DAY = as.numeric(antihistUsers$EVENT_DAY)

# reorder columns
antihistUsers = antihistUsers[,c(1,2,3,4,8,9,10,11,12,13,14,5,6,7)]

# take out purchase events happened before 1995 (121 event ja 14 persons removed)
antihistUsers = antihistUsers[(antihistUsers$SOURCE == "PURCH" & antihistUsers$EVENT_YEAR < 1995),]

# take out R06AE02. Cinnarizine R06AE02 is now days used for nausea. It is not used as antihistamine anymore.
antihistUsers = antihistUsers[(antihistUsers$CODE1 == 'R06AE02'),]

# take out R06AE08. R06AE08 is Levocetirizine but correct ATC code for Levocetirizine in Finnish ATC and WHO ATC is
R06AE09. R06AE08 should not be used for Levocetirizine.
antihistUsers = antihistUsers[(antihistUsers$CODE1 == 'R06AE08'),]

# save to folder
write.table(antihistUsers, file = "/home/ivm/antihistamine/cohorts/antihistUsers.txt", sep = '\t', row.names = FALS
E, col.names = TRUE)
```



Hide

```
# Load cohort from file
antihistUsers = fread("/home/ivm/antihistamine/cohorts/antihistUsers.txt", data.table = FALSE)
```

Hide

```
# How many rows
nrow(antihistUsers)
```

```
[1] 1169125
```

Hide

```
# how many patients
nrow(as.data.frame(table(antihistUsers$FINNGENID)))
```

```
[1] 141737
```

Get frequency table of Antihistamine users

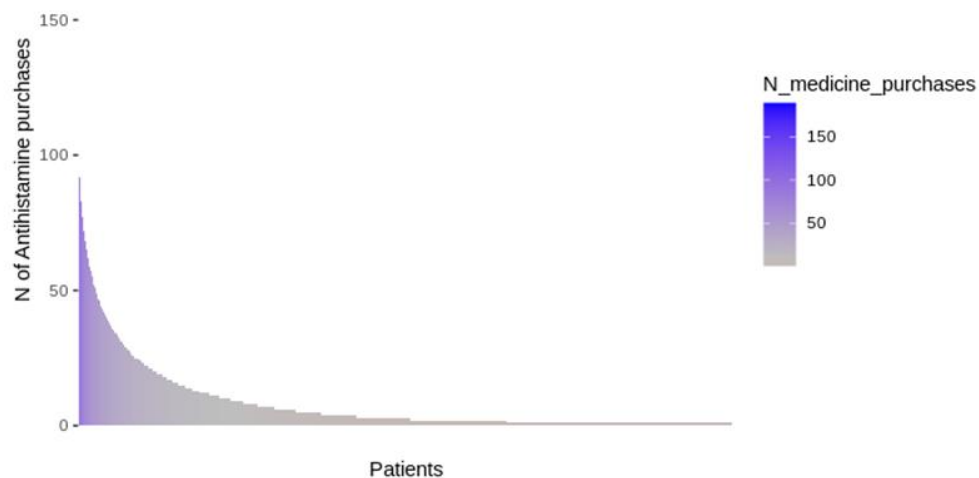
Hide

```
CountsBuyingAntihist = as.data.frame(table(antihistUsers$FINNGENID))
colnames(CountsBuyingAntihist) = c("FINNGENID", "N_medicine_purchases")
```

Plot Antihistamine Users by the number of antihistamine purchases using gradients

Hide

```
ggplot(CountsBuyingAntihist, aes(x = reorder(FINNGENID, -N_medicine_purchases), y = N_medicine_purchases)) +
  geom_col(aes(fill = N_medicine_purchases)) +
  scale_fill_gradient2(low = "red",
                       mid = "grey",
                       high = "blue",
                       #midpoint = median(CountsBuyingAntihist$N_R06A_medicines)) +
                       midpoint = 10) +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank()) +
  xlab("Patients") +
  ylab("N of Antihistamine purchases")
View(antihistUsers)
```



Find age at first purchase of Antihistamine. Include age to the frequency table



Hide

```
DT = data.table(antihistUsers)
AntihistUsers_firstPurchase = DT[DT[, .I[which.min(EVENT_AGE)], by = FINNGENID]$V1]

CountsBuyingAntihist = merge(CountsBuyingAntihist, AntihistUsers_firstPurchase[,c("FINNGENID", "EVENT_AGE")], by = "FINNGENID", all.x = TRUE)
rm(DT)
```

Make age groups 0-9, 10-19, 20-29, ... , 70-89, > 90 and include them as column

Hide

```
CountsBuyingAntihist$AGE_GROUP = findInterval(CountsBuyingAntihist$EVENT_AGE, c(0, 10, 20, 30, 40, 50, 60, 70, 80, 90))
```

Find sex for Antihistamine users. Include sex to the diagnose frequency table

In command line

```
zcat /finngen/library-red/finngen_R7/phenotype_4.0/data/finngen_R7_gt_samples_info.txt.gz | head -n 1
```

```
zcat /finngen/library-red/finngen_R7/phenotype_4.0/data/finngen_R7_gt_samples_info.txt.gz | cut -f 1,3 > SEXofFinnGenIDs.txt
```

Load gender information to RStudio:

Hide

```
# male = 0, female = 1
SEXofFinnGenIDs <- read.table("/home/ivm/SEXofFinnGenIDs.txt", header = TRUE)
```

Hide

```
# replace 0 and 1 with "male" and "female"
v = factor(SEXofFinnGenIDs$SEX)
levels(v) = c("male", "female")
SEXofFinnGenIDs$SEX = v
rm(v)
```

Combine gender to the frequency table

Hide

```
CountsBuyingAntihist = merge(CountsBuyingAntihist, SEXofFinnGenIDs[,c("FINNGENID", "SEX")], by = "FINNGENID", all.x = TRUE)
```

Hide

```
table(CountsBuyingAntihist$AGE_GROUP, CountsBuyingAntihist$SEX)
```

	male	female
1	3093	5544
2	3649	9827
3	4481	12974
4	6849	16115
5	8460	19298
6	9267	17326
7	7088	8929
8	3525	3543
9	875	788
10	30	76

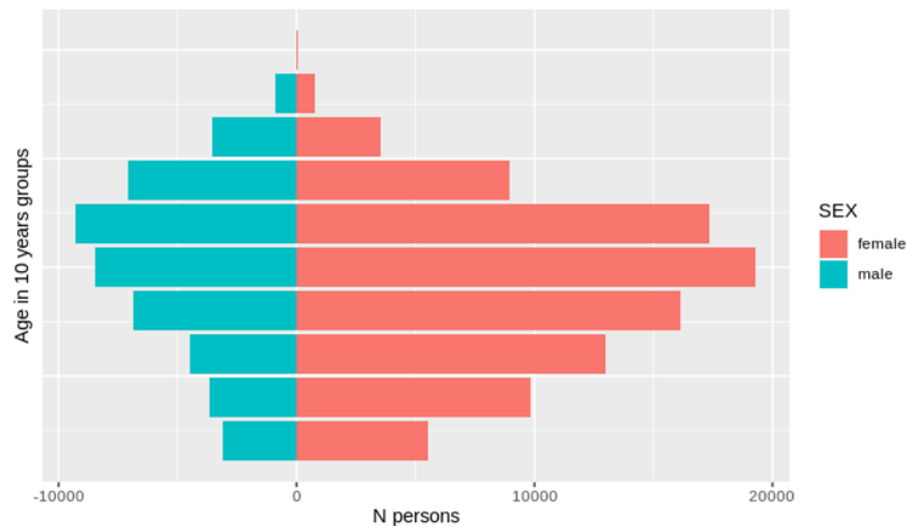
## Plot Population pyramid for Cases group

Hide

```
ggplot(data=CountsBuyingAntihist,aes(x=AGE_GROUP,fill=SEX)) +
  geom_bar(data=subset(CountsBuyingAntihist,SEX=="female")) +
  geom_bar(data=subset(CountsBuyingAntihist,SEX=="male"),aes(y=-.count*(-1))) +
  coord_flip()+
  theme(axis.text.y = element_blank(),axis.ticks.y=element_blank())+

  xlab("Age in 10 years groups")+
  ylab("N persons")+
  ggtitle("Population pyramid for 'Antihistamine users' at age of first purchase")
```

## Population pyramid for 'Antihistamine users' at age of first purchase



## EXPORT list of IDs for custom GWAS command line (cli) tool

Hide

```
# make a cases file for custom GWAS command line tool
antihUsersDF7_IDS = CountsBuyingAntihist[,1]

# save to folder
write.table(antihUsersDF7_IDS, file = "/home/ivm/antihistamine/cohorts/antihUsersDF7_IDS.txt", sep = '\t', quot
e = FALSE, row.names = FALSE, col.names = FALSE)
```

## Defining CONTROL GROUP:

No use of Antihistamine ever - No purchases of R06A - having records after 1st Jan 1995 (because this restriction is also for cases)

Hide

```
# Persons ever used antihistamine
antihistUsersEver <- longitudinal_data_DF7 %>% filter(
  SOURCE == "PURCH" & (str_detect(CODE1, "^R06A") | str_detect(CODE2, "^R06A")))
```

Hide

```
# Longitudinal data for persons never used antihistamines
No_antihistamine = longitudinal_data_DF7[!is.element(longitudinal_data_DF7$FINNGENID,antihistUsersEver$FINNGENI
D),]
```

Hide

```
# take out events happened before 1995
No_antihistamine_1995on_events = No_antihistamine[!(No_antihistamine$EVENT_YEAR < 1995),]
```

Hide

```
# save to folder
fwrite(No_antihistamine, file = "/home/ivm/antihistamine/cohorts/No_antihistamine.txt", sep = '\t', row.names = FALSE, col.names = TRUE)
```

Hide

```
# Load cohort from file
No_antihistamine = fread("/home/ivm/antihistamine/cohorts/No_antihistamine.txt", data.table = FALSE)
```

Hide

```
# remove replicate rows (we need one person only ones in the control 'no antihistamine' table)
No_antihist_pat = No_antihistamine %>% distinct(FINNGENID, .keep_all = TRUE)
```

Hide

```
# How many rows
nrow(No_antihist_pat)
```

```
[1] 178199
```

Hide

```
# how many patients
nrow(as.data.frame(table(No_antihist_pat$FINNGENID)))
```

```
[1] 178199
```

Hide

```
# save to folder
fwrite(No_antihist_pat, file = "/home/ivm/antihistamine/cohorts/No_antihistamine_firstVisitPerPerson.txt", sep = '\t', row.names = FALSE, col.names = TRUE)
```

Hide

```
# Load cohort from file
No_antihist_pat = fread("/home/ivm/antihistamine/cohorts/No_antihistamine_firstVisitPerPerson.txt", data.table = FALSE)
```

Make age groups 0-9, 10-19, 20-29, ... , 70-80, > 90 and include them as a column

Hide

```
No_antihist_pat$AGE_GROUP = findInterval(No_antihist_pat$EVENT_AGE, c(0, 10, 20, 30, 40, 50, 60, 70, 80, 90))
```

Combine gender to the frequency table

Hide

```
No_antihist_pat = merge(No_antihist_pat, SEXofFinnGenIDs[,c("FINNGENID", "SEX")], by = "FINNGENID", all.x = TRUE)
```

Hide

```
table(No_antihist_pat$AGE_GROUP, No_antihist_pat$SEX)
```

```
      male female
1    5269  10212
2    6412   9568
3    8754  12044
4   15548  15347
5   21639  18455
6   18833  12944
7   12717   5341
8    3013   1463
9     210    393
10      7     30
```

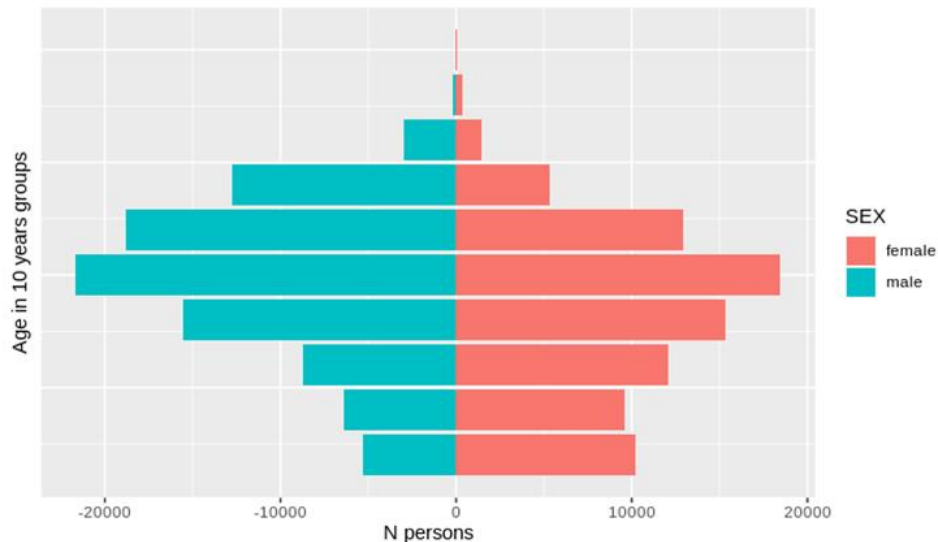
Plot Population pyramid for Control group "Never used antihistamines" at mean age of any event

Hide

```
ggplot(data=No_antihist_pat, aes(x=AGE_GROUP, fill=SEX)) +
  geom_bar(data=subset(No_antihist_pat, SEX=="female")) +
  geom_bar(data=subset(No_antihist_pat, SEX=="male"), aes(y=..count..*(-1))) +
  coord_flip() +
  theme(axis.text.y = element_blank(), axis.ticks.y=element_blank()) +

  xlab("Age in 10 years groups") +
  ylab("N persons") +
  ggtitle("Population pyramid for 'Never used antihistamines' at age of first event")
```

Population pyramid for 'Never used antihistamines' at age of first event



Export list of IDs for custom GWAS command line (cli) tool

Hide

```
# make a control file for custom GWAS command line tool
NoAntihist_IDs = No_antihist_pat[,1]

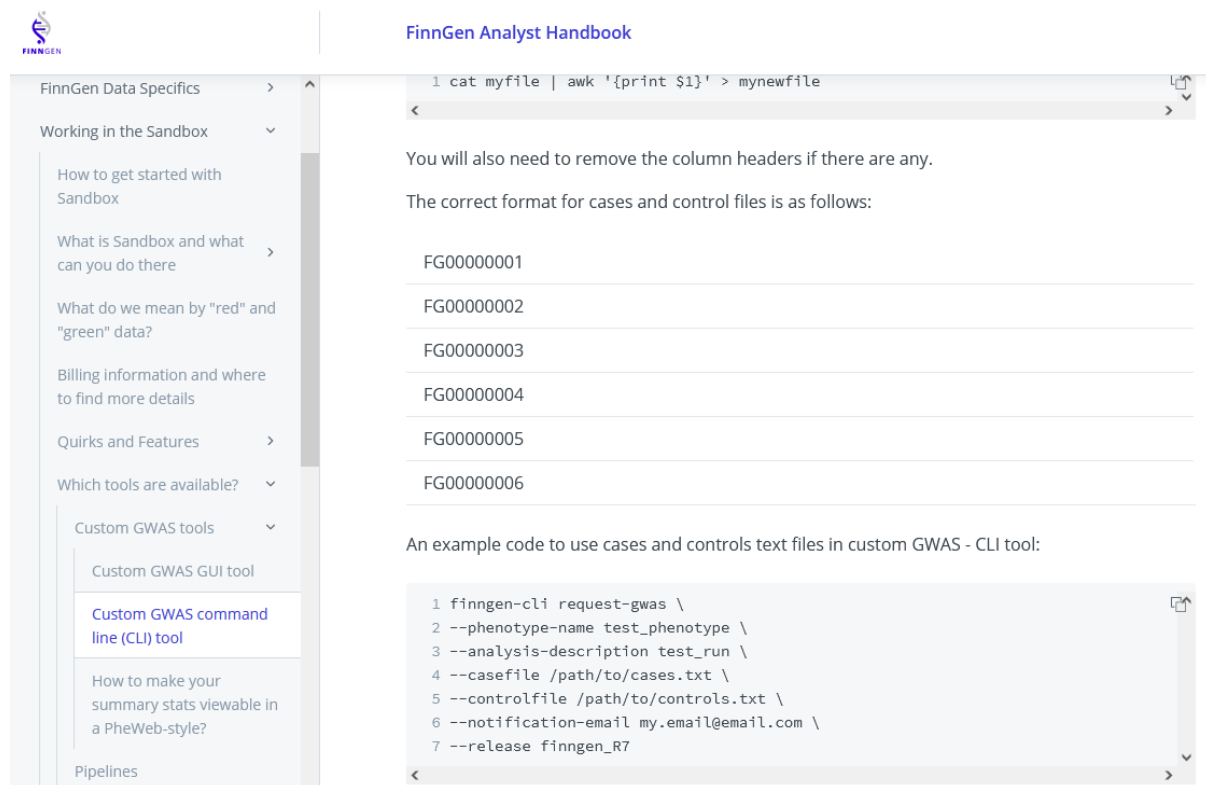
write.table(NoAntihist_IDs, file = "/home/ivm/antihistamine/cohorts/No_antihistamine_forGWAScli.txt", sep = '\t',
  row.names = FALSE, col.names = FALSE, quote = FALSE)
```

**Figure 16.** provides R scripts for cases and controls cohorts building and data visualizations.

### **Starting GWAS with Custom GWAS command line (CLI) tool**

The FinnGen Analyst Handbook [1] describes the three ways the data can be formatted for the custom GWAS command line (CLI) tool (Figure 17). Custom GWAS can be conducted on Atlas cohorts, a list of FinnGen IDs, and from the phenotype file format (Figure 17). In the previous step, the list of IDs was created for the cases and controls cohort in RStudio (see above). Therefore, the second option to conduct GWAS from ID lists is used here.

Figure 18 gives a screen capture from the command needed to start a custom GWAS run on the command line in FinnGen Sandbox. Following the instructions in the FinnGen Analyst Handbook (Figure 17) and using the ID lists created in RStudio (see 7.3.1 Building cohorts with R) phenotype-name, analysis-description, casefile, controlfile and notification e-mail was set to correct values (Figure 18). After successful request Custom GWAS (CLI) tool reports that GWAS analysis request was created successfully and the GWAS run have started (Figure 18).



**Figure 17.** A screen capture from the FinnGen Analyst Handbook section “Custom GWAS command line (CLI) tool” providing detailed instructions on how to conduct a custom GWAS run on the command line. This section in the FinnGen Analyst Handbook is written by the author.

```
ivm@ivm-103662268550497076762:~$ finngen-cli request-gwas --phenotype-name antihistamineUsers --analysis-description antihistamineGWAS --casefile /home/ivm/antihistamine/cohorts/antihUsersDF7_IDS.txt --controlfile /home/ivm/antihistamine/cohorts/No_antihistamine_forGWAScli.txt --notification-email marianna.niemi@tuni.fi --release finngen_R7
11:49:55.524 main ▶ DEBUG Debug logging enabled.
11:49:55.524 goexit ▶ DEBUG main():Running finngen-cli..
11:49:55.656 RunContext ▶ INFO Creating GWAS analysis request
11:50:13.750 RunContext ▶ INFO GWAS analysis request created succesfully
```

**Figure 18.** Screen capture from the commands needed to start a custom GWAS run on the command line using FinnGen custom GWAS command line (CLI) tool.

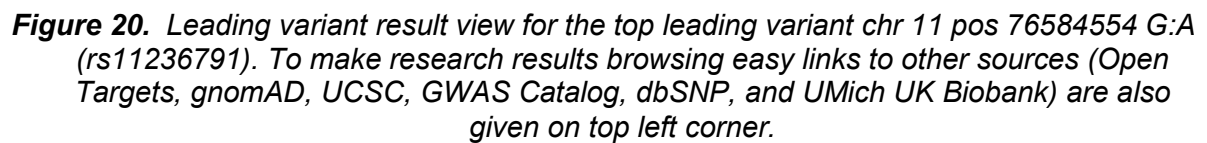
#### 4.2.4 Viewing GWAS results with FinnGen PheWeb tool

Custom GWAS results from custom GWAS (GUI) and custom GWAS (CLI) tools are viewable with FinnGen PheWeb tool [20]. This is the same tool that visualises results also for FinnGen core GWAS analysis from FinnGen endpoints [21]. After 12 months embargo GWAS results will be made publicly available through FinnGen web site [22]. At the time of writing this thesis, the data freeze available to FinnGen researchers is DF8 and DF9 will be available soon. The latest publicly available data freeze is DF6.

FinnGen PheWeb provides a Manhattan plot that is an interactive summary from all GWAS hits for a specific disease endpoint (Figure 19). Hovering mouse over GWAS hits provides information about variants (Figure 19). The level of genome-wide significance is shown with dashed line in Manhattan plot (Figure 19). Information about significantly associated variants is also provided in a table under the Manhattan plot (Figure 19).



**Figure 19.** Screen capture of the FinnGen PheWeb tool front page for Antihistamine user's cohort custom GWAS results. Manhattan plot summarises GWAS hits from custom GWAS results. The table of Lead variants provides chromosome number(chr), position(pos), reference allele(ref), alternative allele(alt), locus, rs-id, nearest gene, consequence, imputation quality score(INFO), enrichment in Finnish population(FIN enrichment), allele frequency(af), allele frequency in cases(af cases), allele frequency in controls(af controls), odds ratio(OR), p-value.



The exploration of the GWAS results tables is made easier by giving links to zoomed views and resources (Figures 19 and 20). From Manhattan plot (Figure 19) the research continues to explore the results on finer scale (Figure 20). The writing of scientific articles can start from the GWAS results. FinnGen also provides core analysis GWAS results that are ready made GWAS from FinnGen endpoints in a similar style as in Figures 19 and 20. The latest ready-made GWAS results are available to FinnGen partner researchers with FinnGen account [21]. After an embargo of 12 months, the ready-made GWAS results are made publicly available to the global scientific community [22].

## 5 CONCLUSIONS

Based on the user interviews and user feedback the FinnGen Analyst Handbook has reached its goal to provide FinnGen researchers useful information, guidance, and efficiency in their research. Users reported that the Analyst Handbook is the first source they seek for help. In most cases users also find the information or guidance they are looking for in the Handbook. If not found in the Handbook, users ask questions in FinnGen community Slack or from FinnGen helpdesk. In many times questions coming to admins through these routes ends up in the FinnGen analyst Handbook where the next user with same question may find the answers. Thus, the Handbook has reached also its second goal to increase knowledge exchange and efficiency within the whole FinnGen project.

Users' feedback is constantly coming in through FinnGen Slack, FinnGen Helpdesk, and through personal contacts with FinnGen staff. The feedback is used to develop and improve FinnGen tools and documentation in the Handbook. The nature of FinnGen Analyst Handbook is that it never be completed and closed but is constantly serving and evolving according to users' needs. The work to make the Analyst Handbook even better continues.



## 6 REFERENCES

- [1] The FinnGen Analyst Handbook (available for FinnGen members - <https://www.finngen.fi/en/members/dashboard>)
- [2] FinnGen project web sites - <https://www.finngen.fi/en> (public site)
- [3] FinnGen Members area - <https://www.finngen.fi/en/members/dashboard> (available with FinnGen account)
- [4] FinnGen Sandbox <https://sandbox.finngen.fi/fg-production-sandbox-<NO>/vm> (NO is the number of Sandbox. Sandbox is available for FinnGen researchers with individual-level data “red data” access)
- [5] Finnish Biobanks - <https://www.biopankki.fi/en/finnish-biobanks/>
- [6] Solita web site - <https://www.solita.fi/en/>
- [7] Risteys web pages for browsing FinnGen endpoints - <https://risteys.finngen.fi/>
- [8] Public FinnGen results - [https://www.finngen.fi/en/access\\_results](https://www.finngen.fi/en/access_results)
- [9] From Alicia Martin at Broad - <https://pan.ukbb.broadinstitute.org/docs/background/index.html>
- [10] Open Targets - <https://genetics-docs.opentargets.org/>
- [11] United Kingdom Biobank, UKBB - [https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/genotyping\\_qc.pdf](https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/genotyping_qc.pdf)
- [12] Ensembl - <https://www.ensembl.org/info/website/index.html>
- [13] Google Documents - <https://docs.google.com/>
- [14] GitBook - <https://www.gitbook.com/>
- [15] GitHub - <https://github.com/>
- [16] Wrike - <https://www.wrike.com/>
- [17] Slack - <https://slack.com/>
- [18] FinnGen documentation pages <https://finngen.gitbook.io/documentation/> (Public documentation about released FinnGen data)

- [19] Darren Murph, Jessica Reeder, Betsy Bula. The importance of a handbook-first approach to documentation <https://about.gitlab.com/company/culture/all-remote/handbook-first-documentation/>
- [20] PheWeb for FinnGen custom GWAS tool results <https://userresults.finngen.fi/>
- [21] PheWeb for the latest FinnGen core GWAS results <https://results.finngen.fi/> (available with FinnGen account)
- [22] PheWeb for publicly available FinnGen core GWAS results <https://r6.finngen.fi/> (older releases after an embargo of 12 months are publicly available)

## 7 APPENDIX

Appendix Table 1. FinnGen Analyst Handbook chapter titles with version number (V) and Publishing date (Pub date) given. Sections written by the author are indicated (X in Author column).

V	Pub Date	FinnGen Analyst Handbook chapter titles	Author
1	15.6.2021	Introduction	
6	14.2.2022	Where to begin...	
6	14.2.2022	I'm a clinician, new to FinnGen, where is the best place for me to start?	X
7	4.4.2022	How do I make a custom endpoint?	
7	4.4.2022	How do I run a GWAS of a phenotype I created myself?	
6	14.2.2022	I'm interested in FinnGen rare variant phenotypes	
2	9.7.2021	Background Concepts	
2	9.7.2021	Basics of Genetics	
2	9.7.2021	Linkage Disequilibrium (LD)	
2	9.7.2021	Genotype Imputation	
2	9.7.2021	Genotype Data Processing and Quality Control (QC)	
2	9.7.2021	GWAS Analysis	
2	9.7.2021	P Values	
6	14.2.2022	Heritability and genetic correlations	
2	9.7.2021	Finemapping	
2	9.7.2021	Colocalization	
2	9.7.2021	Using Polygenic Risk Scores	
2	9.7.2021	PheWAS analysis	
2	9.7.2021	Longitudinal Data Analysis	
2	9.7.2021	Introduction to Atlas	
2	9.7.2021	GWAS Association to Biological Function	
2	9.7.2021	Genetic Data Resources outside FinnGen	
3	1.10.2021	Getting Started with Unix	
2	9.7.2021	Getting Started with R	
1	15.6.2021	FinnGen Data Specifics	
2	9.7.2021	FinnGen Data Freezes and Releases	
1	15.6.2021	Data and data access	
1	15.6.2021	Structure of the FinnGen project	
1	15.6.2021	What is unique about Finnish healthcare and health data?	
1	15.6.2021	Finland as a Population Bottleneck	
1	15.6.2021	What kind of questions can I ask of FinnGen data?	
1	15.6.2021	Data safety and protection	
1	15.6.2021	How to request an account	
2	9.7.2021	Analysis proposals	
2	9.7.2021	What is a FinnGen analysis proposal and when do I need to submit one?	
2	9.7.2021	How do I submit an analysis proposal?	

2	9.7.2021	How are analysis proposals handled?	
2	9.7.2021	How do I submit a bespoke analysis proposal?	
2	9.7.2021	What is the difference between FinnGen analysis proposals and FinnGen bespoke analyses?	
2	9.7.2021	Genotype data	
2	9.7.2021	Genotype Arrays Used	
7	4.4.2022	Legacy cohorts and chips	
2	9.7.2021	Imputation Panel	
3	1.10.2021	Sisu v4 reference panel	
2	9.7.2021	Sisu v3 reference panel	
2	9.7.2021	Genome build used in FinnGen	
2	9.7.2021	Genotype Data Processing Flow	
2	9.7.2021	Genotype Files in Sandbox	
6	14.2.2022	Imputation data file	
6	14.2.2022	bgen file	
6	14.2.2022	Chip data file	
6	14.2.2022	Imputed STR genotypes	
6	14.2.2022	Genotype plink data	
6	14.2.2022	Imputed HLA alleles	
6	14.2.2022	PCA data	
6	14.2.2022	Kinship data	
2	9.7.2021	Analysis covariate file	
6	14.2.2022	PRS data	
6	14.2.2022	GRM data	
6	14.2.2022	Prune data	
2	9.7.2021	Finnish Health Registries and Medical Coding	
2	9.7.2021	Finnish health registries	
5	10.12.2021	Register data pre-processing	
2	9.7.2021	International and Finnish Health Code Sets	
3	1.10.2021	More information on health code sets	
3	1.10.2021	Mapping FinnGen Longitudinal Data to the OMOP-Common Data Model	X
2	9.7.2021	Register code translation files	
2	9.7.2021	Phenotype Files in Sandbox	
2	9.7.2021	Detailed longitudinal data	
6	14.2.2022	What are combination codes and how they are separated in detailed longitudinal data?	
5	10.12.2021	Registers in the detailed longitudinal data	
2	9.7.2021	Endpoint and endpoint longitudinal data	
2	9.7.2021	Minimum phenotype and longitudinal data	
7	4.4.2022	Extraction of FinnGen minimum data set information per biobank	
7	4.4.2022	DNA isolation protocols per biobank	
2	9.7.2021	Cohort data	
2	9.7.2021	Other registry data files in Sandbox	
2	9.7.2021	Endpoints	
2	9.7.2021	Location of FinnGen Endpoint and Control Description Files	
7	4.4.2022	What's new in DF9 endpoints	

5	10.12.2021	What's new in DF8 endpoints
5	10.12.2021	Interpretation of Endpoint Definition file
3	1.10.2021	Location of Endpoint Quality Control Report
2	9.7.2021	Creating a User-defined Endpoint(s)
2	9.7.2021	Requesting a User-defined Endpoint to be included in Core Analysis
2	9.7.2021	How to use PheWeb
2	9.7.2021	How to use Ristey as an Endpoint Browser
2	9.7.2021	Data Masking/Blurring of Visit Dates
6	14.2.2022	Complete follow-up time of the FinnGen registries – primary endpoint data
6	14.2.2022	Survival analysis using the truncated endpoint file – secondary endpoint data
2	9.7.2021	Publishing FinnGen results
2	9.7.2021	Citing FinnGen
2	9.7.2021	The 1-year "Exclusivity Period" Policy
2	9.7.2021	How to send an Application to the Scientific Committee
2	9.7.2021	List of Publications using FinnGen Data
2	9.7.2021	Public Result Releases
2	9.7.2021	Green Library Data (aggregate data)
2	9.7.2021	What is "Green" Data?
2	9.7.2021	Accessing Green Data
2	9.7.2021	Other analyses available
2	9.7.2021	Colocalizations in FinnGen
2	9.7.2021	Autoreporting – information on overlaps
2	9.7.2021	Index of Autoreporting variables
2	9.7.2021	HLA
3	1.10.2021	Meta-analysis of FinnGen with UK and Estonian Biobanks
6	14.2.2022	Core analysis results files
6	14.2.2022	Genotype cluster plots format
6	14.2.2022	GWAS results format
6	14.2.2022	Finemapping results format
6	14.2.2022	Colocalization results format
6	14.2.2022	Autoreporting results format
6	14.2.2022	UKBB-FinnGen meta-analysis file formats
6	14.2.2022	Estonian BB-UKBB-FinnGen meta-analysis file formats
6	14.2.2022	Pairwise endpoint genetic correlation format
6	14.2.2022	Heritabilities
6	14.2.2022	Coding variant associations format
6	14.2.2022	Chip GWAS
1	15.6.2021	Working in the Sandbox
2	9.7.2021	How to get started with Sandbox
1	15.6.2021	What is Sandbox and what can you do there
1	15.6.2021	What sort of work can you do in the sandbox that you can't do anywhere else?
1	15.6.2021	What do we mean by "red" and "green" data?
1	15.6.2021	Billing information and where to find more details
2	9.7.2021	Quirks and Features

2	9.7.2021	Navigating the sandbox	X
2	9.7.2021	Copying and pasting in and out of your IVM	
2	9.7.2021	How to report issues from within the Sandbox	X
2	9.7.2021	Sharing individual-level data within the Sandbox	
2	9.7.2021	How to download results from your IVM	
2	9.7.2021	Keyboard combinations	
2	9.7.2021	Running analyses in your IVM vs. Pipelines	
2	9.7.2021	Timeouts and saving your work (backups, github)	
2	9.7.2021	How to upload to your own IVM via /finngen/green	
3	1.10.2021	How to install a R package into Sandbox?	X
3	1.10.2021	How to install R packages with many dependencies	
3	1.10.2021	How to install a Python package into Sandbox	X
6	14.2.2022	How to install GNU Debian package	X
5	10.12.2021	Sandbox IVM tool request handling policy	X
4	1.11.2021	Docker images	X
2	9.7.2021	How to get a new Docker image to Sandbox	X
3	1.10.2021	How to mount data into Docker container image	X
3	1.10.2021	Containers available to Sandbox	X
3	1.10.2021	Containers with user customized tool sets	X
4	1.11.2021	How to write a Docker file	X
3	1.10.2021	Python Virtual Environment in Sandbox	X
2	9.7.2021	How to shut down your IVM	X
2	9.7.2021	Which tools are available?	X
2	9.7.2021	Custom GWAS tools	X
2	9.7.2021	Custom GWAS GUI tool	X
2	9.7.2021	Custom GWAS command line (CLI) tool	X
3	1.10.2021	How to make your summary stats viewable in a PheWeb-style?	
2	9.7.2021	Pipelines	X
2	9.7.2021	Unix tools	
2	9.7.2021	R libraries	
2	9.7.2021	Atlas	X
2	9.7.2021	BigQuery (relational database)	X
2	9.7.2021	Genotype Browser	
2	9.7.2021	Jupyter	X
2	9.7.2021	How to check which programs are available in Sandbox	X
2	9.7.2021	Red Library Data (individual data)	
2	9.7.2021	Working with Genotype Data	
3	1.10.2021	Genotype Browser how to	
3	1.10.2021	Cluster Plots	
6	14.2.2022	Install ClusterPlot viewer V3C	
6	14.2.2022	Rare Variant Calling in V3C	
3	1.10.2021	Create map of allele	
3	1.10.2021	Genotypes from VCF files	
3	1.10.2021	Variant PheWas	
6	14.2.2022	Interpreting rare-variant analysis results	
2	9.7.2021	Working with Phenotype Data	

3	1.10.2021	Variant PheWas	
2	9.7.2021	How to define a cohort in Atlas	X
3	1.10.2021	Standard and non-standard medical codes in Atlas	X
3	1.10.2021	Filtering by clinical registries in Atlas	X
4	1.11.2021	Selecting configuration in Atlas	X
5	10.12.2021	Defining exit rules for a cohort in Atlas	X
5	10.12.2021	How to export FinnGen IDs from Atlas	X
6	14.2.2022	Cohort Characterizations in Atlas	X
7	4.4.2022	Interpreting the results of Feature Analysis in ATLAS	X
7	4.4.2022	Improving cohorts using Cohort Characterizations tool	X
5	10.12.2021	How to select controls for your cases	
3	1.10.2021	Using the R libraries to look at Phenotype data	
5	10.12.2021	How to check case counts from the data	
2	9.7.2021	Creating your own user-defined endpoint	
2	9.7.2021	Running analyses in Sandbox	
2	9.7.2021	How to use the Pipelines tool	X
3	1.10.2021	Pipelines is based on Cromwell and WDL	
2	9.7.2021	How to submit a pipeline from the command line (finngen-cli)	X
6	14.2.2022	How to run genome-wide association studies (GWAS)	
2	9.7.2021	How to run GWAS using REGENIE	
6	14.2.2022	Running quantitative GWAS with REGENIE	X
2	9.7.2021	How to run GWAS using SAIGE	X
6	14.2.2022	How to run GWAS using plink2 (for unrelated individuals only)	
2	9.7.2021	How to run GWAS using GATE (survival models)	
6	14.2.2022	How to run finemapping pipeline	
2	9.7.2021	How to run PRS	
6	14.2.2022	How to calculate PRS weights for FinnGen data	
4	1.11.2021	Sandbox path and pipeline mappings	X
7	4.4.2022	If your pipeline job fails	X
7	4.4.2022	Tips on how to find a pipeline job ID	X
6	14.2.2022	Managing memory in Sandbox and data filtering tips	X
2	9.7.2021	FAQ	
2	9.7.2021	FinnGen access and accounts	
2	9.7.2021	Do I need "red" or "green" data access?	
2	9.7.2021	Where do I apply for data access?	
2	9.7.2021	If I already have green data access, how do I apply for red?	
2	9.7.2021	How do I enable two-factor authentication?	
2	9.7.2021	What if I can't access my FinnGen account?	
2	9.7.2021	How to reset account credentials	
2	9.7.2021	What to do if you suspect your account has been compromised	
2	9.7.2021	Can't access your smartphone for 2FA?	
2	9.7.2021	How do I access the FinnGen Sharepoint and Members area?	
3	1.10.2021	FinnGen All Sharepoint (e-duuni) site - how to log in for the first time?	
6	14.2.2022	How can I view existing analysis proposals?	
2	9.7.2021	Can I join the FinnGen Slack?	
2	9.7.2021	FinnGen data	

2	9.7.2021	I think I found a mistake in the longitudinal data?	
2	9.7.2021	What are the field/column names in FinnGen?	
2	9.7.2021	What covariates are used in FinnGen's core GWAS analyses?	
2	9.7.2021	Does FinnGen have lab results available?	
2	9.7.2021	Does FinnGen have family and relatedness information available?	
2	9.7.2021	Where can I find a list of unrelated individuals in FinnGen?	
2	9.7.2021	When moving from BCOR to .txt files, what does the column called "correlation" mean?	
2	9.7.2021	What's the difference between phenotype_2.0 and phenotype_3.0?	
2	9.7.2021	Is there really no participant birth year data?	
2	9.7.2021	How do I calculate time between events?	
2	9.7.2021	Can I select only the columns needed for my analysis to import into RStudio?	
2	9.7.2021	How do I get patient IDs for analysis?	
2	9.7.2021	Can I download all pairwise LD data across the genome at once?	
2	9.7.2021	Where can I find	
2	9.7.2021	COVID association results?	
2	9.7.2021	Users' Meeting slides?	
2	9.7.2021	A list of what coding variants are enriched in Finland?	
2	9.7.2021	A comprehensive list of key file locations in FinnGen?	
2	9.7.2021	Medical code translations?	
2	9.7.2021	PheWeb	
3	1.10.2021	What are QQ and Manhattan plots?	
2	9.7.2021	How do I get access to PheWeb?	
2	9.7.2021	Can I have the fine-mapping results available in PheWeb as flat files?	
2	9.7.2021	Do the autoreports report the 95% or 99% credible set?	
7	4.4.2022	Volcano plots with LAVAA	
2	9.7.2021	Registries	
2	9.7.2021	What do KELA reimbursement codes map to?	
2	9.7.2021	What's the cutoff date for FinnGen data?	
2	9.7.2021	Sandbox	
2	9.7.2021	Where can I find tutorials on Sandbox?	
2	9.7.2021	Where can I read Sandbox documentation?	
2	9.7.2021	Can I copy text from Sandbox to my computer?	
2	9.7.2021	How do I get my own code files into Sandbox?	
2	9.7.2021	Is there a way to share individual level data between different Sandbox users?	
2	9.7.2021	Is there a sun grid engine for running long scripts?	
3	1.10.2021	How to clear browser cache after sandbox update	X
2	9.7.2021	How do I increase the window resolution on my IVM?	
3	1.10.2021	How can I view pdf, jpg and HTML files?	
7	4.4.2022	How to apply SES sandbox access	
2	9.7.2021	Atlas	
5	10.12.2021	Can I use FinnGen endpoints like medical codes in Atlas?	X
2	9.7.2021	Risteys	
2	9.7.2021	How do I access Risteys?	
5	10.12.2021	Why is the case number dropping after the "Check pre-conditions, main-only, mode, ICD version" step?	



2	9.7.2021	Endpoints	
2	9.7.2021	Where do I find the most recent list of FinnGen endpoints?	
2	9.7.2021	What does it mean when an endpoint has “mode” at the end?	
2	9.7.2021	What scenario would cause an NA (missing data) entry rather than a zero?	
2	9.7.2021	Does it mean anything when a value is written as \$!\$ instead of NA?	
2	9.7.2021	Why is there an inconsistency between ICD10 code J84.1 (IPF) and J84.112?	
2	9.7.2021	Are FinnGen's hypothyroid cases similarly defined?	
2	9.7.2021	How are control endpoints calculated?	
2	9.7.2021	Can I get a list of FinnGen IDs by control group for my endpoint?	
2	9.7.2021	Should I expect all control-endpoint IDs to be excluded from normal-endpoint ones?	
2	9.7.2021	What does Level C mean in the endpoints data table?	
2	9.7.2021	What does the SUBSET_COV field show?	
2	9.7.2021	Why is there a "K." prefix on some endpoints?	
6	14.2.2022	Why there are fewer endpoints going from R5 (N = 2,925) to R8 (N = 2,202)?	
6	14.2.2022	I found BL_AGE after FU_END_AGE in the endpoint data, how is it possible?	
6	14.2.2022	Why individuals who are not dead have death age in endpoint data?	
6	14.2.2022	I found EVENT_AGE after FU_END_AGE in endpoint data, how is it possible?	
2	9.7.2021	Pipelines	
2	9.7.2021	Are there example SAIGE pipelines?	
2	9.7.2021	How do I apply finemapping to my SAIGE results?	
2	9.7.2021	Publications	
2	9.7.2021	Using public data	
2	9.7.2021	How do I cite an analysis that used publicly available FinnGen data?	
2	9.7.2021	Using private data (for partners)	
2	9.7.2021	How do I cite FinnGen in analyses that use private data?	
2	9.7.2021	Is there any standard FinnGen text I can use as a reference?	
2	9.7.2021	For the biobanks	
2	9.7.2021	How to apply for data return	
2	9.7.2021	Where to ask for software you'd like to see in Sandbox	
3	1.10.2021	Release Notes	X