GENOTYPE IMPUTATION OF KUOPIO BREAST CANCER PROJECT DATA

Irene Pöllänen

Pro gradu -tutkielma

Itä-Suomen yliopisto

Luonnontieteiden ja metsätieteiden tiedekunta

Biotieteet, Biokemia

2017

ITÄ-SUOMEN YLIOPISTO
Luonnontieteiden ja metsätieteiden tiedekunta, biotieteet, biokemia
PÖLLÄNEN, IRENE: Kuopio Breast Cancer Project -datan genotyyppi-imputaatio
Pro gradu –tutkielma
Pro gradu –tutkielman ohjaajat: Arto Mannermaa ja Arja Tervahauta
Kesäkuu 2017

Genotyyppi-imputaatio on menetelmä määrittämättömien genotyyppien päättelemiseksi. SNP-sirulla genotyypitettyä SNP-joukkoa voidaan imputoimalla laajentaa sisältämään sirun SNP-joukkoon kuuluttamia SNP:itä. Menetelmällä voidaan muun muassa lisätä tehokkuutta assosiaatiotutkimuksiin, joissa voidaan etsitä mm. sellaisia geneettisiä variaatiota, joilla on assosiaatio esimerkiksi rintasyöpään.

IMPUTE2 on yksi useista tarjolla olevista ja yleisesti käytetyistä haplotyyppien vaiheistus- ja genotyyppi-imputaatio-ohjelmista. Genotyyppi-imputaatio perustuu haplotyyppien hyödyntämiseen referenssitietokannoista.

Kuopio Breast Cancer Projekti (KBCP) suoritettiin vuosina 1990-1995. Tutkimukseen kutsuttiin osallistumaan naisia, jotka kävivät Kuopion yliopistollisessa sairaalassa rintaoireiden takia. Osalta tutkimukseen osallistuneista naisista tehtiin genotyypitys iCOGS SNP-sirulla. Tämä data on arvokasta materiaalia rintasyövän geneettisten tekijöiden määrittämiseksi. Tässä Pro gradu -tutkielmassa käytetty KBCP-genotyyppidata koostuu rintasyöpään sairastuneista yksilöistä sekä referenssiyksilöistä.

Tässä tutkielmassa, imputoitiin 696 yksilön kromosomi 22 KBCP-genotyyppidatasta IMPUTE2-ohjelmalla käyttäen 1000G-haplotyyppidataa referenssinä. Imputaation tuloksena iCOGS-sirulla genotyypitetty 3 597 SNP:n joukko laajentui1 106 696 SNP:n estimoiduilla genotyyppien todennäköisyyksillä. Suurin osa SNP:istä imputoitiin 1.0 todennäköisyysarvolla. Imputaation konkordanssin keskiarvo oli noin 95% ja jokaisessa imputaatiolohkossa yksilöiden korrelaation neliö oli suurempi kuin 0.6. Nämä imputaatiostatistiikka-arvot ylittivät toivotusti IMPUTE2:n mainitsemat muissa tutkimuksissa aiemmin käytetyt imputaation laadun arvioinnin arvot valideille imputaatioille. Koko KBCP iCOGS genomin genotyyppidatan imputaatio kestäisi arviolta noin 1.5 kuukautta UEF:n Bioinformatiikkakeskuksen klusterissa ajettuna ja mahdollisesti alle päivän esimerkiksi CSC:n laskennallisella kapasiteetilla.

Imputaatiossa on tärkeää käyttää mahdollisimman laajaa referenssijoukkoa. Esimerkiksi suomalaista genotyyppidataa imputoidessa todennäköisesti tärkeämpää on, että referenssijoukko on mahdollisimman suuri, kuin että se olisi puhtaasti populaatiospesifistä.

Avainsanat: genotyyppi-imputaatio, genotyyppi, rintasyöpä, IMPUTE2, bioinformatiikka

UNIVERSITY OF EASTERN FINLAND
Department of Environmental and Biological Sciences, bioscience, biochemistry
PÖLLÄNEN, IRENE: Genotype imputation of Kuopio Breast Cancer Project Data
Master's Thesis
Supervisors of the Master's Thesis: Arto Mannermaa and Arja Tervahauta
June 2017

Genotype imputation is a method for inferring unobserved genotypes. The original set of genotyped SNPs within an array is extended with initially-ungenotyped SNPs to provide more power to, e.g., association studies for finding genetic variants for association with a certain trait, e.g., breast cancer.

IMPUTE2 is one of the several available programs for genotype imputation. It is a haplotype phasing and genotype imputation program. Genotype imputation is based on using haplotypes from reference data sets.

Kuopio Breast Cancer Project (KBCP) is a study conducted in 1990-1995. Finnish women with breast symptoms entering Kuopio University Hospital were invited to take part to the study. A subset of the individuals that took part to the study is genotyped with a iCOGS SNP array. This data is valuable material for detecting genetic factors for breast cancer. The KBCP genotype data used in this Pro gradu -thesis work consists of 445 individuals with breast cancer and 251 reference individuals.

In this thesis, the imputation of chromosome 22 from KBCP-genotype data was performed using IMPUTE2 software and 1000G haplotype data as a reference. The output was the original 3 597 genotyped SNPs of chromosome 22 from KBCP -data extended with the genotype probabilities of 1 106 697 initially-ungenotyped SNPs imputed from 1000G. Majority of SNPs were imputed with the probability value of 1.0. The average concordance was around 95 % and the squared correlation of the individuals from imputed chunks were over 0.6. These imputation quality statistics are above the cut-offs used in previous experiments, as mentioned by IMPUTE2, for valid imputation result. The estimated imputation time for the whole KBCP iCOGS genotype genome data would be around 1.5 months with the UEF Bioinformatics Center's cluster and possibly less than a day using e.g. the larger computational capacity provided by CSC's.

It is crucial to use as large a reference data set as possible when performing imputations. It is likely that it is most important to have as large as possible a reference data set instead of just limiting the reference data into a population specific set.

Keywords: genotype imputation, genotype, breast cancer, IMPUTE2, bioinformatics

Abbreviations and Definitions

CNV	*Copy number variation*; form of genetic variation where the number of copies of a certain DNA sequence varies between individuals
DNA	*Deoxyribonucleic acid*; polynucleotide carrier of genetic information consisting of adenine, cytosine, guanine and thymine
GWAS	*Genome wide association study*; approach for identifying genetic variants associated with a certain trait
KBCP	*Kuopio Breast Cancer Project*; data set containing genotyped SNP array data gathered from a group of Finnish females with breast cancer, also the set contains reference data
SNP	*Single Nucleotide Polymorphism*; nucleotide position in the genome where at least 2 alternative nucleotides are common in a population (rare allele frequency > 1%)

Index

Attachments

1 Introduction

Genotype imputation is a method for increasing the number of SNPs that can be tested in association studies. It increases the number of defined genotypes from a SNP array by estimating ungenotyped SNPs from reference data and adding those estimates into the dataset. Various imputation tools currently exist and several reference datasets are available for use in imputation purposes. Moreover, imputation has already been an important step in many genome-wide association studies for searching previously undefined susceptibility traits from genetic data.

Kuopio Breast Cancer Project (KBCP) is a study conducted in 1990-1995. Finnish women with breast symptoms entering Kuopio University Hospital were invited to take part to the study. A subset of the individuals that took part to the study is genotyped with a iCOGS SNP array. The KBCP genotype data used in this Pro gradu -thesis work consists of 445 individuals with breast cancer and 251 reference individuals. This data potentially contains still undefined inherited susceptibility factors for breast cancer, and imputation has the potential to increase the odds for detecting them.

In this thesis, the steps of the imputation process and its principles are explained. As the experimental part of this thesis, the imputation of KBCP chr22 genotype data was performed using IMPUTE2 software and 1000G haplotype data as a reference. The result was an imputed chr22 where the original set of 3 597 genotyped SNPs was increased with genotype probabilities to 1 106 697 SNPs. Majority of SNPs were imputed with the probability value of 1.0. The average concordance was around 95 % and the individual level the squared correlation measure calculated from imputed chunks were over 0.6. These imputation quality statistics are above the cut-offs used in previous experiments, as mentioned by IMPUTE2, for valid imputation result. The estimated imputation time for the whole KBCP iCOGS genotype genome data would be around 1.5 months with the UEF Bioinformatics Center's cluster and possibly less than a day if larger computational capacity offered by e.g. CSC is used.

In addition to this introduction, this thesis contains five sections. Section 2 is the literature review part of the thesis, containing the definitions of the terms and concepts related to this topic. Following this, the research objectives and hypothesis are represented in Section 3. The materials and methods are in section 4. As this research is conducted fully in silico, section 4 focuses

heavily on the code used in this experiment. In section 5, the results and output of the imputation are presented. Finally, conclusions of this thesis are presented in chapter 6.

2 Genotype imputation of breast cancer SNP array data

Genetic variance can be investigated with, e.g., *genome-wide association studies* (GWAS). GWAS studies can be used in, e.g., finding susceptibility factors to breast cancer, which is the most common form of cancer in Finnish women. As a preliminary step of GWAS studies, genotype imputation can be used as means to increase the amount of SNP array study data by filling in untyped SNPs from reference datasets[1].

In this section of the thesis an overview of central concepts concerning the topic the this is given. The first section covers breast cancer (2.1) and the second is about the human genome and Finnish inheritance (2.2). This is followed by a short section about SNPs and genetic variance (2.3). Last section (2.4) concentrates on genotype imputation, imputation software, reference datasets, quality measures and the role of imputation in association studies.

2.1 Breast cancer

Breast cancer is the most common cancer in Finnish women. In 2014, 5008 Finnish women in total were diagnosed with breast cancer[2]. Fortunately, increased understanding about the disease, advances in cancer treatments during recent decades, and early detection of the lesion have had beneficial effects on the prognosis.

Breast cancer, like cancer in general, is a result of abnormal growth of a tissue. This uncontrolled growth is due to several compromised mechanisms that lead to disastrous traits, including self-sufficiency in growth signals and insensitivity to anti-growth signals[3].

Breast cancer can vary in its form and expression of genes. This has led to different classification schemes of the disease, e.g. to ER+ (estrogen-receptor-positive) and ER-. Recognizing the type and traits of the lesion enables more fine-tuned treatment, e.g. hormonal therapy to ER+ patients. Naturally, estrogen-receptor-positivity is just one of the many possible traits of this complex disease[4].

It is estimated that the already known breast cancer susceptibly genes explain only 20-25 % of the Finnish families with high rates of breast cancer[5]. Therefore, much is still unknown about the underlying genetics that cause susceptibly to breast cancer.

2.2 Human genome and Finnish inheritance

The human genome is about $3*10^9$ bases (3 Gb) in length and consists mostly of linear nuclear DNA. The nuclear genome is divided into 22 autosomal chromosome pairs and two sex chromosomes (XX or XY)[6]. In addition, the human genome consists of a very small fraction of circular mitochondrial DNA. In 2014, it was estimated that there are about 19 000 protein-coding genes in the human genome[7]. These genes are unevenly divided between and within the chromosomes and 37 of the genes are located in the mitochondrial DNA[8].

Every human individual is genetically unique (excluding identical twins). This uniqueness is due to a process of recombination and mutations. In recombination, homologous chromosomes exchange DNA segments during meiosis. This event leads to offspring that differ in their combination of traits from those found in either parent. However, the exchange rates between genes are not completely independent. Genes close to each other tend to stay together in recombination than genes more far apart. And naturally, phenotypically expressible traits that are encoded to neighboring genes should therefore also occur more frequently together in the offspring.

Mutations are random changes in DNA. A single nucleotide mutation occurs when a single nucleotide is changed into another, eliminated or inserted into the DNA sequence as an additional nucleotide. The result may have no effect or it can lead to, e.g., non-functioning gene product. Mutations can also lead to beneficial effects providing more favorable traits to cope with the changing environment. Thus, mutations are an important and natural part of evolution.

The susceptibility to different diseases, like cancer, can be inherited encoded in the DNA in families. The Finnish population is known to originate from a small group of isolated ancestors[5]. This has led to a *founder effect*, i.e., the founders' genetic information, including ancient mutations, are enriched in the Finnish population. Some genetic traits among Finns are more and some less frequent compared to the rest of the World's population, causing the Finnish disease heritage to differentiate from that of other populations.

2.3 SNPs and genetic variance

As previously mentioned, the genetic code of every human individual is unique to some extent (excluding identical twins). Humans share on average 99.9 % of their genetic code with other members of the species. However, considering that the human genome is 3 billion base pairs in

length – even 0.1 % creates a significant amount of variation between individuals. In human DNA variations ranges from single nucleotide changes to changes in the number of chromosomes.

However, all nucleotide positions do not vary in the same rate in a population. Only about one nucleotide in 300 in the human genome tends to exist in different forms[8]. These *polymorphic* nucleotides are SNPs (*single nucleotide polymorphisms*) (see Figure 1). A SNP is defined as a nucleotide position in the genome where at least 2 alternative nucleotides are common in a population (rare allele frequency > 1%). It is numerically the most abundant genetic variation. The different possible versions of the SNP are called *alleles*. Typically, a SNP has two alternative alleles and the typical ratios of the alleles can vary in different populations. The more frequent allele is known as *major allele* and the less frequent as *minor allele*. The human chromosomes are diploid, i.e., there are two copies of each autosomal chromosome. Thus, every cell has two alleles of each SNP. This allele pair can consist of two different alleles (*heterozygous*) or two same alleles (*homozygous*) and is called a *genotype*.

Moreover, the SNP may influence the phenotype, or it might be neutral. And, some SNPs have higher level of significance to the association with a trait than others. In complex diseases, e.g., breast cancer, multiple SNPs shows association with susceptibility to the disease.[8]

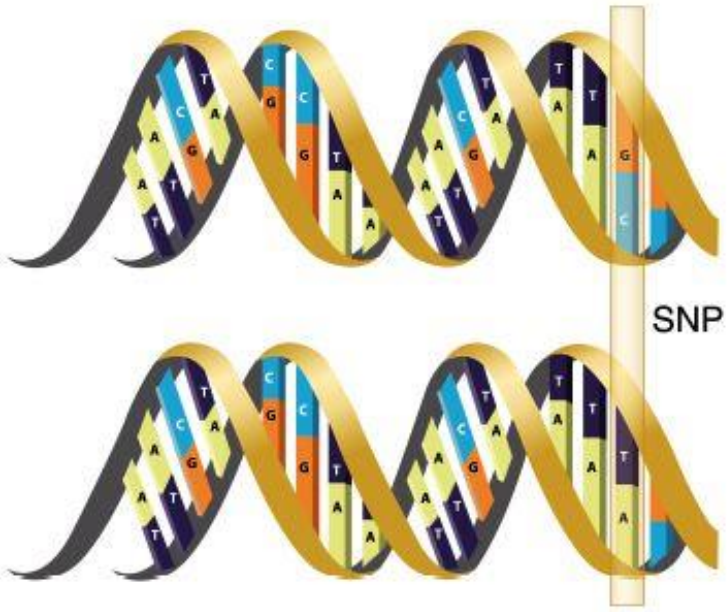


Figure 1. An illustration of a SNP. Most nucleotides in a population are quite constant. Only 1 of 300 is a so-called SNP – a nucleotide that varies in at least 1 % in a population. In this example figure a SNP is highlighted from the two complementary DNA strands while the other nucleotide positions are identical having the same nucleotide. (image source: [9])

Since genes close to each other tend to stay together in recombination than genes more far apart, this applies naturally also to SNPs. The blocks of genetic code, linked *loci* (unique locations of genes), that tend move on to offspring from one chromosome are called *haplotypes*. The loci in haplotypes are said to be in linkage disequilibrium.[8]

2.4 Genotype imputation

SNP arrays are DNA microarrays that are used in genotyping SNPs. They can be used to define patient's diseases susceptibilities or suitability of drug therapies. Typically, SNP arrays genotype a set of certain SNPs, e.g., iCOGS arrays that were used to genotype genetic variants related to three hormone related cancers: breast, ovary and prostate (total of 211 155 SNPs).

Even though SNP arrays genotype only a fraction of all the possible SNPs in a population, the undefined SNPs can still be estimated to some extent using a procedure known as imputation.

The underlying principle in imputation is the use of haplotypes from reference haplotypes (Figure 2). Imputation can be performed on the whole genome for, e.g., genome wide association studies, or just on small part for a fine-mapping study.

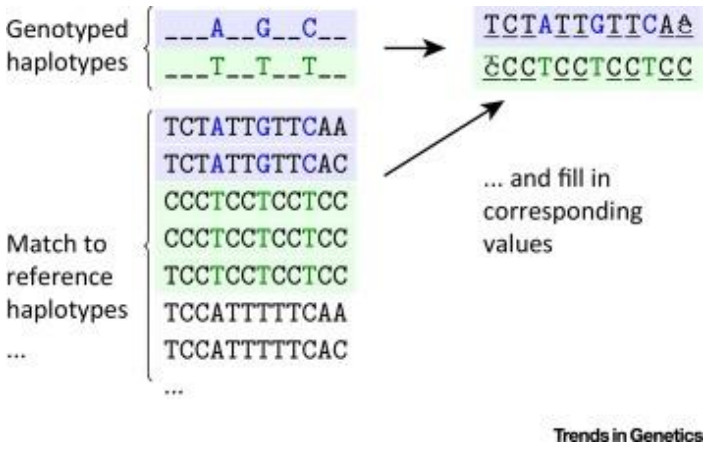


Figure 2. Genotype imputation. Using reference haplotypes, the undefined SNPs are added into a set of genotyped SNPs. (image source: [10])

Imputation usually involves two parts. The first part is haplotype phasing of unphased diploid data, which means that data in diploid form needs to be separated into haploid form to perform haplotype matching. This is done by using known haplotypes from a reference and aligning. The next part then involves imputation of haploid data by aligning the defined SNPs to reference haplotypes and filling the holes in the study data with the alleles from the matching reference haplotype.

2.4.1 Software

Genotype imputation has been commonly used as a technique for several years and several different imputation software have been developed. IMPUTE is a genotype imputation and haplotype phasing program developed by Howie et al in 2009[11]. Its latest release is called IMPUTE2 v2.3.2 and it was released in December 2014[12]. It is a command line based program for Linux. And, it is based on a Hidden Markov model and Markov Chain Monte Carlo framework. As output IMPUTE gives genotype probabilities for the imputed SNPs. The genotype probabilities are between 0-1 but the sum of the possible genotype probabilities for one SNP does not necessarily sum to 1. IMPUTE2 is also the imputation tool used in the experimental part of this thesis.

Other genotype imputation programs are e.g. BEAGLE, SHAPEIT and minimac[13]. The genotype imputation tools are typically designed to run in a Linux environment and do not provide versions for other operating systems, e.g. Windows. Experiments that compare the performance of these genotype imputation programs have been performed[14], and IMPUTE2 has performed well in these comparisons.

2.4.2 Reference data sets

A crucial premise of imputation is the availability of a large amount of reference data. It is most straightforward to use a single reference panel, e.g., 1000G. IMPUTE2 provides a straight link to the 1000G data set in a ready to use format[15]. Naturally, the reference data set can also be a custom-made combination of various reference data sets.

It is generally considered that the more the better when it comes to the amount of genomic reference data. Therefore, the reference data sets are continuously under enhancement, with more data continuously generated and added to the existing data pool. But bigger reference datasets also mean that the imputation becomes computationally more demanding and the imputation process can become costly with increased resource needs (e.g. in terms of purchased computation time or plain electricity).

2.4.3 Quality measures

Quality measures describing imputation accuracy either compare imputation results to known true genotype data or estimate the quality without known true genotypes. In a typical scenario dataset is imputed without the ground truth (true genotypes). Concordance rate and squared correlation R^2 are quality measures commonly relied upon in such a situation. IMPUTE2 provides concordance rates and squared correlation as default from imputation experiments. Both squared correlation R^2 and concordance values are formed by masking one variant at a time from the input data (genotype file) and then imputing those variants back using nearby study variants and reference data. [16]

2.4.4 Genotype imputation for association studies

The imputed data itself does not translate to novel scientific knowledge or medical discoveries without further studies. The end goal of imputation is generally to perform association studies. In such a study the set of genomic variants are tested if they have an association with a certain

trait, e.g., susceptibility to breast cancer. Imputing genotyped SNP array data is a more cost effective for analysing variants than next-generation sequencing in a large study cohort.

For example, a study was published in 2015 where genotype imputation was used to generate genotypes for more than 11 million SNPs for a GWAS were 15 new susceptibly loci for breast cancer were identified[17]. In another study, imputed array data was used identify hundreds of variants associated with age at menarche and the supporting causal risk for breast cancer[18].

3 Research objectives

The aim of this study was to impute KBCP genotype data using IMPUTE2 and 1000G reference dataset. The original KBCP genotype data was collected with a custom Illumina iSelect genotyping array, iCOGS[19] and the total number was 199 961 genotyped SNPs from 696 individuals. The imputation work in this thesis, was limited to imputing only genotyped SNPs from chromosome 22 which is the second shortest somatic chromosome.

For imputation to be possible at all, the study data and reference data should have some overlapping genotyped SNPs. The bigger the number of shared SNPs the better the imputation result should be. Also, the more reference data there is, the better the quality of imputations should be.

The first assumption in this study was that there would be enough overlapping SNPs in chromosome 22 between the study and reference data. The second assumption was that the imputation would be possible for this dataset in general. A logical question would then be how well the output would enhance the GWAS and would this lead to new findings. The last assumption though, is not tested in this thesis. Therefore, statements of the benefits and usefulness of the imputed data in further analysis will not be made in this thesis.

The purpose of this study was mainly to investigate how the practical process of imputation is performed. After establishing the pipeline and principles, it can then be assessed if genotype imputation is something that could be considered to experiment with another set up with the same data. One possible such research set up would be using only Finnish reference data. Additionally, the imputation time of the whole genome is estimated based on the imputation time of chromosome 22.

The result of this imputing experiment would be the probabilities of initially-ungenotyped genetic variants added to the original set of genotyped SNPs from the iCOGS array and the quality measures of the imputation. Another part of this study is to review the literature and research already done with genotype imputation in general. The motivation of this study is also to provide statements about the reasonability of a possible future study in which only Finnish reference data would be used.

4 Material and Methods

In this section the experimental part of the work performed for this thesis is covered. First the Kuopio Breast Cancer Project -data utilized in this experiment is described in section 4.1. Then in section 4.2 the steps of the data pre-processing are explained. This section has heavy emphasis on the scripts that were used to pre-process the KBCP genotype data to a format suitable for IMPUTE2. Finally, in section 4.3 a description of the actual imputation performed with IMPUTE2 is provided.

4.1 Kuopio Breast Cancer Project SNP genotype data

Prior to this study, Kuopio Breast Cancer Project (KBCP) genotype data was collected with a custom Illumina iSelect genotyping array, iCOGS[19], from 696 individuals. The total number of genotyped SNPs was 199 961 in each array.

KBCP is a study conducted in 1990-1995. Finnish women with breast symptoms entering Kuopio University Hospital were invited to take part to the study. A subset of the individuals that took part to the study was genotyped with a iCOGS SNP array. The KBCP genotype data used in this Pro gradu -thesis work consists of 445 individuals with breast cancer and 251 reference individuals.

The original KBCP data for this study consisted of two files: *a genotype file* (Table 1) containing the genotypes for all 199 961 SNPs of the individuals and *a SNP location file* (Table 2) containing the position information for the SNPs. The data was in csv-format.

Table 1. Definitions of the variables in the genotype file.

variable name	range	description	number
BCAC_ID	[‘KBCP-270’, ‘KBCP-185’…,’KBCP-623’,’KBCP-629’]	patient index	696
SNP	[200047’,’200050’, …’SNP49’,’SNP68’]	SNP allele pair reads of chromosome pairs	199 961
genotype	['AA', 'GG', 'AC', 'AG', 'TT', 'CC', 'AT', 'CG', 'II', '00', 'DD']	genotype for every SNP of each patient	696*199961

Table 2. Definitions of the attributes linked to each SNP in the SNP location file

Variable name	Range	Description
SNP_index	1 to 211155	the list of SNPs is indexed from 1 to 211155, latter presenting the overall number of defined SNPs in the raw dataset
Illumina_ SNP_Name	['200047' '200050' '200251' ..., 'SNP127' 'SNP49' 'SNP68']	SNP Illumina names
LocalName	['rs3755048' 'rs2276637' 'rs1043502' ..., 'c10_pos135201254' 'c7_pos99203930' 'c7_pos99205761']	SNP local names
New_rs_number	[nan 'rs115397167' 'rs116295078' ..., 'rs71490244' 'rs55800129' 'rs11552449']	new rs number for each SNP
Chromosome	['2' '1' '6' '10' '11' '12' '13' '14' '15' '16' '17' '18' '19' '20' '21' '22' 'X' 'Y' 'MT' '3' '4' '5' '7' '8' '9' 'XY']	the chromosome number where the SNP is located
Position_Build36	[219793146 219797929 20850335 ..., 135201254 99203930 99205761]	chromosome location of the SNP in human reference genome build36
Position_Build37	[220084902 220089685 20977748 ..., 135351264 99365994 99367825]	chromosome location of the SNP in human reference genome build37
ReferenceStrand	['+' '-' 'u']	the reference genome strand info

4.2 Data pre-processing

Data pre-processing was performed both to become more familiar with the KBCP data in general, and to extract and convert the study material into a suitable input format for IMPUTE2. Python[20] was chosen as the scripting language due to its prevalent use in the field of bioinformatics and the availability of mature libraries for data analysis. Combined with libraries, e.g. pandas[21] (Python Data Analysis Library) and NumPy[22], it provides a modern and robust toolset for scientific data manipulation and analysis.

To facilitate reproducibility, the intermediate results in the pre-processing pipeline were stored in a human readable format, e.g., as csv-files. Thus, the correctness of each step could then be assessed straightforwardly from the stored output files.

The original SNP location file contained more SNPs than the KBCP genotype file. Therefore, as a first pre-processing step the location file was trimmed by removing those SNPs that were not included in the KBCP genotype file. The dataset resulting from this was thus an intersection of the SNPs from the location and genotype file. This was done with the following python script (`extract_KBCP_locations_info.py`):

```
import pandas as pd

SNP_data=pd.read_csv('SNP_genotype_file.csv',sep='\t', low_memory=False)

#get SNP IDs to a list from the header

SNP_list = SNP_data.columns.values.tolist()

#remove 'BCAC_ID' from the list since it not an SNP

SNP_list.remove('BCAC_ID')

#open the output csv

wfile = open("KBCP_iCOGS_SNP_locs_info.csv", "w")

SNP_info = pd.read_csv('SNP_location_file.csv', sep=',', low_memory=False)

for i in range(0,SNP_info.shape[0]):

  if (SNP_info.iloc[i,1] in SNP_list):

     row = ''
```

```
        SNPs = SNP_info.iloc[i,:]
        for j in range(0,SNPs.shape[0]):
                row = row + str(SNPs.iloc[j]) + '\t'
        row = row + '\n'
        wfile.write(row)
    wfile.close()
```

The imputation to be performed in this thesis was limited to imputation of chromosome 22 only. Therefore, the genotype data from chromosome 22 was to be extracted. First the SNPs from chromosome 22 were identified from the SNP location file and extracted with the following python script (`extract_chr22_locations_info.py`):

```
import pandas as pd
SNP_data = pd.read_csv('KBCP_iCOGS_SNP_locs_info.csv', sep =',',
low_memory= False)
wfile = open("chr22_iCOGS_SNP_locs.csv", "w")
for i in range(0,SNP_data.shape[0]):
          if (SNP_data.iloc[i,4] == '22'):
                      row = ''
                      SNPs = SNP_data.iloc[i,1:]
                      for j in range(0,SNPs.shape[0]-1):
                      row = row + str(SNPs.iloc[j]) + '\t'
                      row = row + '\n'
                      wfile.write(row)
wfile.close()
```

The output contained the location info for 3597 SNPs. After this extraction, the Illumina_ SNP_Name:s from the extracted file were used as IDs to identify the corresponding genotyped

SNPs from the genotype file for extracting these genotypes with the following python script (`extract_chr22_genotypes.py`):

```
import pandas as pd

Illumina_SNPs = pd.read_csv('chr22_iCOGS_SNP_locs.csv', sep ='\t', low_memory=False, header=None)

#get the Illumina SNP ID names

SNP_IDs = Illumina_SNPs.iloc[:,0]

#reading the KBCP genotypes

KBCP_genotypes = pd.read_csv('KBCP_snps.txt', sep ='\t', low_memory=False)

data = KBCP_genotypes[pd.Series(['BCAC_ID']).append( SNP_IDs)]

data.to_csv(path_or_buf='chr22_iCOGS_SNP_genotypes.csv', index=False)
```

Now the output contained the genotypes for the 3597 SNPs from chromosome 22. The different genotypes and their frequencies from this output are listed in Table 3. In addition to genotypes with defined nucleic acid allele pairs, there were 3 356 genotypes with missing information (marked as 00) and 696 defined as deletions (marked as DD). Inspection of the data showed that the 00 reads were scattered throughout the data and the deletions were all from one SNP (rs17882761).

Table 3. The genotypes from chromosome 22 SNPs in the study dataset to be imputed.

genotype	description	number
GG	guanine-guanine	807 179
AA	adenine-adenine	713 711
AG	adenine-guanine	588 758
CC	cytosine-cytosine	177 768
AC	adenine-cytosine	114 292
CG	cytosine-guanine	48 270
TT	thymine-thymine	27 846
AT	adenine-thymine	20 940
00	no read - no read	3 356
DD	deletion-deletion	696

For the next step, the genotypes had to be converted to *genotype file format.* It is the format in which genotypes of a study cohort to be imputed have to be as input for IMPUTE2. [23] gives guidelines and an example about genotype file format as follows:

"The genotype file stores data on a one-line-per-SNP format. The first 5 entries of each line should be the SNP ID, RS ID of the SNP, base-pair position of the SNP, the allele coded A and the allele coded B. The SNP ID can be used to denote the chromosome number of each SNP. The next three numbers on the line should be the probabilities of the three genotypes AA, AB and BB at the SNP for the first individual in the cohort. The next three numbers should be the genotype probabilities for the second individual in the cohort. The next three numbers are for the third individual and so on. The order of individuals in the genotype file should match the order of the individuals in the sample file (see below). Also, the probabilities need not sum to 1

to allow for the possibility of a NULL genotype call. This format allows for genotype uncertainty. This genotype file format is the same as that produced by the genotype calling algorithm CHIAMO.

Example

Suppose you want to create a genotype for 2 individuals at 5 SNPs whose genotypes are

SNP 1 : AA AA
SNP 2 : GG GT
SNP 3 : CC CT
SNP 4 : CT CT
SNP 5 : AG GG

The correct genotype file would be

SNP1 rs1 1000 A C 1 0 0 1 0 0
SNP2 rs2 2000 G T 1 0 0 0 1 0
SNP3 rs3 3000 C T 1 0 0 0 1 0
SNP4 rs4 4000 C T 0 1 0 0 1 0
SNP5 rs5 5000 A G 0 1 0 0 0 1

So, at SNP3 the two alleles are C and T so the set of 3 probabilities for each individual correspond to the genotypes CC, CT and TT respectively.

***Note** : columns 2 and 3 (that contain the RS ID and base-pair position of the SNPs are set arbitrarily in this example."*

Following these instructions, the genotypes for the 3597 SNPs from chromosome 22 were converted to genotype file format with the following python script (`convert_to_genotype_file_format.py`):

```
import pandas as pd

SNP_data = pd.read_csv('chr22_iCOGS_SNP_genotypes.csv', sep =',',
low_memory=False)

#getting the full list without the patient ID:s

SNP_list=SNP_data.iloc[:,1:]

#get SNP names/IDs

SNP_names = SNP_data.columns.tolist()

SNP_info= pd.read_csv('chr22_iCOGS_SNP_locs.csv', sep = '\t',
low_memory=False, header=None)

wfile = open("KBCP_chr22_genotypes_in_gff.csv", "w")

location = 'NULL'

SNP_count = 0

for i in range(0,SNP_list.shape[1]):

  SNP_ID = SNP_names[i+1]

  #get unique SNPs for forming genotypes

  SNPs = SNP_list.iloc[:,i].unique().tolist()

  if '00' in SNPs:

      SNPs.remove('00')

  #concatenate all alleles

  sentence = ""

  for word in SNPs:

      sentence += "" + word

  #eg. AAGGAG to AG:

  allelepair = ''.join(set(sentence))
```

```
#chromosomal location
SNP_location = SNP_info.iloc[i,5]
SNP_name = SNP_info.iloc[i,0]
SNP_count += 1
if SNP_ID == SNP_name:
    location = SNP_location
genotypes = " "
alleles = ""
#if SNP is heterozygous in the data
if len(allelepair)==2:
    allele1 = allelepair[0]
    allele2 = allelepair[1]
    homo_pair1 = allele1 + allele1
    hetero_pair1 = allele1 + allele2
    hetero_pair2 = allele2 + allele1
    homo_pair2 = allele2 + allele2
    for j in range(0, SNP_list.shape[0]):
        KBCP_allele_pair = SNP_data.iloc[j,i+1]
        if KBCP_allele_pair == homo_pair1:
            genotype = '1 0 0 '
        elif (KBCP_allele_pair == hetero_pair1 or KBCP_allele_pair ==
        hetero_pair2):
            genotype = '0 1 0 '
        elif KBCP_allele_pair == homo_pair2:
```

```
                genotype = '0 0 1 '
            else:
                genotype = '0 0 0 '
            genotypes = genotypes + genotype
        alleles = allele1 + " " + allele2
    else:
        allele1 = allelepair[0]
        KPCP_allele_pair = allele1 + allele1
        for j in range(0, SNP_list.shape[0]):
            genotype = '1 0 0 '
            genotypes = genotypes + genotype
        alleles = allele1 + " " + allele1
        row = ('SNP%i'%(SNP_count)+" "+ SNP_names[i+1] + " " +
        str(location) + " " + alleles + genotypes)
    row = row + '\n'
    wfile.write(row)
wfile.close()
```

Some of the SNPs in the genotype file did not contain any heterozygous forms. Thus, the minor allele was not known. There was no clear answer given in the instructions provided by creators of IMPUTE2 for how situations like this should be handled. For this imputation, when there was only A as major and no minor, the genotype file format was decided to contain the following:

SNP1 rs1000 1000 AA 0 0 1

After this step, SNPs were not yet in ascending order in the output. In IMPUTE2 SNPs must be sorted by their position. Therefore, the output was next organized with the following python script (`order_genotype_file.py`):

```
import pandas as pd

SNP_data = pd.read_csv('KBCP_chr22_genotypes_in_gff.csv', sep =' ',
low_memory=False, header = None)

id_colum = SNP_data.ix[:,0].copy()

sorted_SNP_data = SNP_data.sort_values(2)

sorted_SNP_data.ix[:,0] = id_colum.values

sorted_SNP_data.to_csv('KBCP_chr_genotypes_gff.csv', index = None, header
= None, sep = ' ')
```

After this pre-processing step, the genotype file for imputation was ready. Next, a strand orientation file was to be extracted from the original study data. IMPUTE2 assumes that strand orientation for the input genotypes is + (forward) unless otherwise noted. However, the genotypes of KBCP SNPs are not always from the + strand. Therefore, the strand orientation info needed to be extracted to create a suitable strand orientation file. IMPUTE2[12] instructs the following about strand orientation file:

"File showing the strand orientation of the SNP allele codings in the ***-g*** *file, relative to a fixed reference point. Each SNP occupies one line, and the file should have two columns: (i) the base pair position of the SNP and (ii) the strand orientation ('+' or '-') of the alleles in the genotype file; the columns should be separated by a single space.*

The ordering of the SNPs in this file does not matter (by contrast to the ***-g*** *file, which must be sorted by SNP position), and it is okay if some SNPs in the strand file are not present in the genotype file (e.g., due to filtering)."*

In accordance with the instructions, the strand orientation file was built with the following python script (`extract_chr22_SNP_reference_strand.py`):

```
import pandas as pd

SNP_loc_info = pd.read_csv('chr22_iCOGS_SNP_locs.csv', sep ='\t',
low_memory=False, header = None)

#get Position_Build37 and ReferenceStrand columns

SNP_strand_info = SNP_loc_info.iloc[:,5:7]
```

```
SNP_strand_info.to_csv(path_or_buf='chr22_iCOGS_SNP_strand.csv',    index=
False, sep=';', header = None)
```

The script provided as output a strand orientation file (`chr22_iCOGS_SNP_strand.csv`) - containing SNP locations with 1758 – (reverse) strands, and 1820 + strands and 19 u (undefined) strands. The undefined strand orientations were then manually queried via Kaviar[24] (see Figure 4) to produce strand estimates (see Table 4). Those estimates were then used to replace the undefined labels from the strand orientation file.

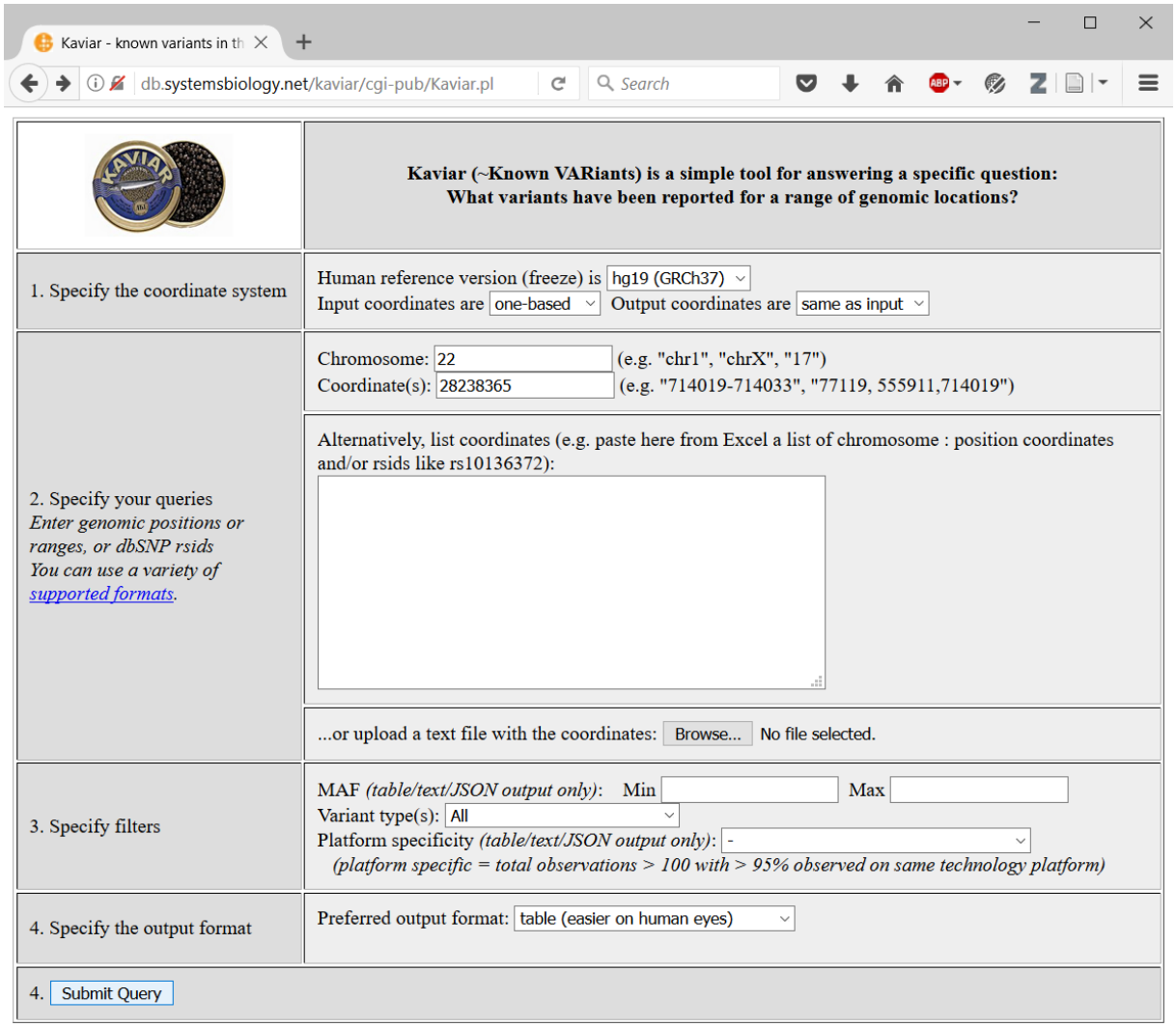


Figure 4. Kaviar Genomic Variant Database provides a tool to query SNP information. This figure is a snap shot of the Kaviar UI from the website.

Table 4. SNPs from chromosome 22 with originally undefined strand orientation with strand orientation estimates made with Kaviar queries. Strand estimates marked with * could not be clearly estimated,therefore those are merely guesses. (IMPUTE2 checks the strand orientation during the run, so even guesses should not be severe problems for the output.)

SNP_ index	IlluminaID	Position _Build37	orig. alleles	orig. major	KAVIAR major	KAVIAR alleles +strand	strand orientation estimate (-/+)
21855	rs1062731	28248840	AG	G	C	CT	- (-strand major: G)
31364	rs11090516	28244191	AG	G	C	CT	- (-strand major: G)
46867	rs12170011	28238365	AC	C	G	GT	- (-strand major: C)
60358	rs13054858	25424579	AC	C	C	CA	+ (+strand major: C)
85391	rs17432784	17196300	AG	G	C	CT	- (-strand major: G)
92599	rs1883472	28224350	AG	G	C	CT	- (-strand major: G)
95925	rs1997739	28250172	AG	A	C	CT	- (-strand major: G)
96066	rs2001194	33056782	CG	C	G	GC	-* (-strand major: C)
115980	rs2740982	33069689	AG	G	C	CT	- (-strand major: G)
122538	rs3016111	17183103	AG	G	C	CT	- (-strand major: G)
127081	rs361995	17164478	AG	G	C	CT	- (-strand major: G)
145865	rs4821083	33056341	AG	A	T	TC	- (-strand major: A)
152549	rs5748636	17275394	AG	G	A	AG	+* (+strand major: A)
152629	rs5752043	25416023	AG	G	T	TC	- (-strand major: A)
152630	rs5752044	25419524	AG	G	C	CT	- (-strand major: G)
154632	rs5996821	25449213	AC	A	T	TG	- (-strand major: A)
177648	rs7286373	25427646	AG	G	G	GA	+ (+strand major: G)
178065	rs730647	28250810	AG	G	C	CT	- (-strand major: G)
205445	rs9613536	28241138	AG	G	G	GA	+ (+strand major: G)

4.3 Genotype imputation of chromosome 22

IMPUTE2 was installed on the University of Eastern Finland Bioinformatics Center's servers[25]. The application was downloaded and installed from the official IMPUTE2 website[12] and then extracted using gzip. 1000G was also downloaded from [15] and used as the reference data in the imputation. 1000G, a.k.a., Thousand Genomes is a publicly available reference data set that has been commonly used in imputations. It is also the reference data that IMPUTE2 uses in its examples. The downloaded reference 1000G dataset is based on NCBI build 37 and contains sequence data for 2504 samples from [26].

IMPUTE2 provides examples and guidelines for using the IMPUTE2 program, e.g., required arguments, input file options and output file options (for additional details, see[12]). Since IMPUTE2 recommends imputation of a whole chromosome to be performed in smaller chunks, a Slurm[27] Workload Manager batch script was written to process the individual chunks separately. Slurm is an open source workload manager and it enables partitioning of a submitted job to smaller chunks that the servers connected to the cluster can then independently run on their own time. The server on which IMPUTE2 was installed for this thesis work had access to a Slurm cluster and enabled submitting the batch script to Slurm using the sbatch command. The imputation of the whole chromosome 22 was done in chunks of 3 000 000 base pairs.

The imputation of KBCP genotypes with unobserved genotypes was done with the following script (IMP_chr22_SBATCH.sbatch):

```
#!/bin/bash -l

#SBATCH -J IMP

#SBATCH -n 10

#SBATCH -t 30-00:00:00

#SBATCH -o /home/users/ipollane/imputointi/IMP_chr22_sbatch_output.txt

#SBATCH -e /home/users/ipollane/imputointi/IMP_chr22_sbatch_error_log.txt


mkdir /home/users/ipollane/imputointi/imputed_chr22
```

```
mkdir /home/users/ipollane/imputointi/imputed_chr22_times


TASKID=${SLURM_ARRAY_TASK_ID}


echo "Start time: $(date)" >> /home/users/ipollane/imputointi/imputed_chr22_times/slurm_execution_times_${SLURM_ARRAY_TASK_ID}.log


./impute2 \
 -m ./1000GP_Phase3/genetic_map_chr22_combined_b37.txt \
 -h ./1000GP_Phase3/1000GP_Phase3_chr22.hap\
 -l ./1000GP_Phase3/1000GP_Phase3_chr22.legend\
 -g ./Imputation_test1/KBCP_gff_chr22.csv \
 -strand_g ./Imputation_test1/chr22_iCOGS_SNP_strand.csv\
 -int $((16000000+(TASKID-1)*3000000)) $((16000000+(TASKID)*3000000)) \
 -Ne 20000 \
 -o /home/users/ipollane/imputointi/imputed_chr22/impute_batch_${SLURM_ARRAY_TASK_ID}.impute2
echo "End time: $(date)" >> /home/users/ipollane/imputointi/imputed_chr22_times/slurm_execution_times_${SLURM_ARRAY_TASK_ID}.log
```

The reference data inputs contained: the fine-scale recombination map for the region to be analysed; `genetic_map_chr22_combined_b37.txt`, a file of known haplotypes; `1000GP_Phase3_chr22.hap`, and its legend file with information about the SNPs; `1000GP_Phase3_chr22.legend`. The input files contained the KBCP genotypes in the genotype file format (`KBCP_gff_chr22.csv`) and the strand orientation file (`chr22_iCOGS_SNP_strand.csv`) which were generated with the scripts shown earlier in the data pre-processing section.

The flag `-int` defines the chromosome location interval for each batch. Here the chromosome interval is set to 3 000 000. The flag `-Ne` gives the effective size of population from which the dataset is sampled. IMPUTE2 suggests setting `-Ne` to 20000 in general, and therefore that recommendation was used in this study as well.

4.4 Analysis of the imputed genotypes

The impute result files contain the imputed SNPs and their genotype probabilities. The imputation result for a single initially-ungenotyped SNP comprises of the probabilities of the three possible genotypes (a triplet). For example, the first imputed SNP from location 16 050 075 of the first study individual had the following distribution:

AA 0.997 AG 0.003 GG 0

We can see from the result that the imputation result suggests that the genotype AA would be the most likely genotype with high confidence. Genotypes AG or GG are not considered likely at all. In general, we could think that a genotype imputation like in this situation with one of the likelihoods being substantially higher than the others would be more meaningful than the following situation:

AA 0.33 AC 0.33 CC 0.33

In this latter situation, all the genotypes are as likely and there would not be a single genotype for the SNP that could be considered as the most likely of this individual.

To see how many of the SNPs were given a high value to one of its possible genotype in a triplet, the maximum values were extracted from the impute result file and the frequencies between 0.0-1.0 in intervals of 0.0-0.1 were extracted in Figure 5 with the following script (`extract_max_probability.py`):

```
import pandas as pd

import numpy as np

import pickle



def mapping_function(index):
```

```
   return np.floor((index -5)/3)

def getFrequencies(data):

   imputed_indexes = data[data[0].str.contains('---')].index.tolist()

   largest  =  data.iloc[imputed_indexes,5:-1].groupby(by=mapping_function,
axis=1).max();

   largest = largest.stack()

   return  pd.cut(largest,  np.arange(0.0,1.1,0.1)  ).value_counts(),  larg-
est.mean()

total_counts = None;

overall_mean = 0

chunks_processed = 0;

for i in range(1,13):

   chunksize = 10 ** 4

   for chunk in pd.read_csv('S:\\IMPUTATION PROJECT\imputed_chr22

   \impute_batch_% i.impute2'%(i), sep =' ', low_memory=False,

   header = None,chunksize=chunksize):

      chunk_counts,mean = getFrequencies(chunk)

            if total_counts is None:

               total_counts = chunk_counts

            else:

               total_counts = total_counts.add(chunk_counts, fill_value = 0)

            chunks_processed += 1

            overall_mean += mean

overall_mean = overall_mean / chunks_processed

with open('frequences','w') as f:
```

```
   pickle.dump([total_counts, overall_mean], f, pickle.HIGHEST_PROTOCOL)
```

After this extraction, the results were plotted in a histogram with the following script (`plot_histogram.py`) (Figure 5):

```
import numpy as np

import matplotlib.pyplot as plt

import pickle


def autolabel(rects):

    for rect in rects:

        height = rect.get_height()

        ax.text(rect.get_x() + rect.get_width()/2., 1.01*height,

                '%d' % int(height),

                ha='center', va='bottom')

data = pickle.load( open('frequences', 'rb'))

data = data[0]

histogram = data[0];

fig, ax = plt.subplots()

rects = ax.bar(np.arange(0,data.shape[0]),data)

ax.set_xticks(np.arange(0,data.shape[0]+1))

ax.set_xticklabels(np.arange(0.9,1.01,0.01))

autolabel(rects)

plt.show()

print(np.sum(data))
```

5 Results

Since the imputation of chromosome 22 was done in 12 chunks, naturally the resulting imputed chromosome consisted of 12 sets of imputation results. After the completion of all individual runs, the chunks were combined with the following command in Linux:

```
cat impute_batch_1.impute2 impute_batch_2.impute2 impute_batch_3.impute2 impute_batch_4.impute2 impute_batch_5.impute2 impute_batch_6.impute2 impute_batch_7.impute2 impute_batch_8.impute2 impute_batch_9.impute2 impute_batch_10.impute2 impute_batch_11.impute2 impute_batch_12.impute2 > chr22_All.impute2
```

From each single chunk run, IMPUTE2 produces the following output files: summary, impute, info, info by sample and warnings.

The summary file is a log file that records a summary of the console (screen) output. The summary file from the first chunk is given at the end of this thesis as an example (see Appendix 1). The summary file also contains the concordance table, which can be used to examine if there were problems with the imputation.

5.1 Quality measures

IMPUTE2 performs internal cross-validation and provides squared correlation R^2 and concordance values as quality measures for imputation performance evaluation utilizing the input data. Both squared correlation R^2 and concordance values are formed by masking one variant at a time from the input data (genotype file) and then imputing those variants back using nearby study variants and reference data.

The cross-validation results are shown in the concordance table and listed in the info-files outputted by the program. Low concordance between the imputed and input genotypes may indicate problems in the imputation. Also, the overall concordance should generally be around 95 % (the number on the upper right corner of the table).

The overall concordances are listed in Table 5, along with the SNP statistics of each chunk.

Table 5. SNP statistics of each chunk from the imputation of chromosome 22. The table shows the overall concordance and the distribution of SNPs detected from input and reference data. The last column shows the imputation time for each chunk.

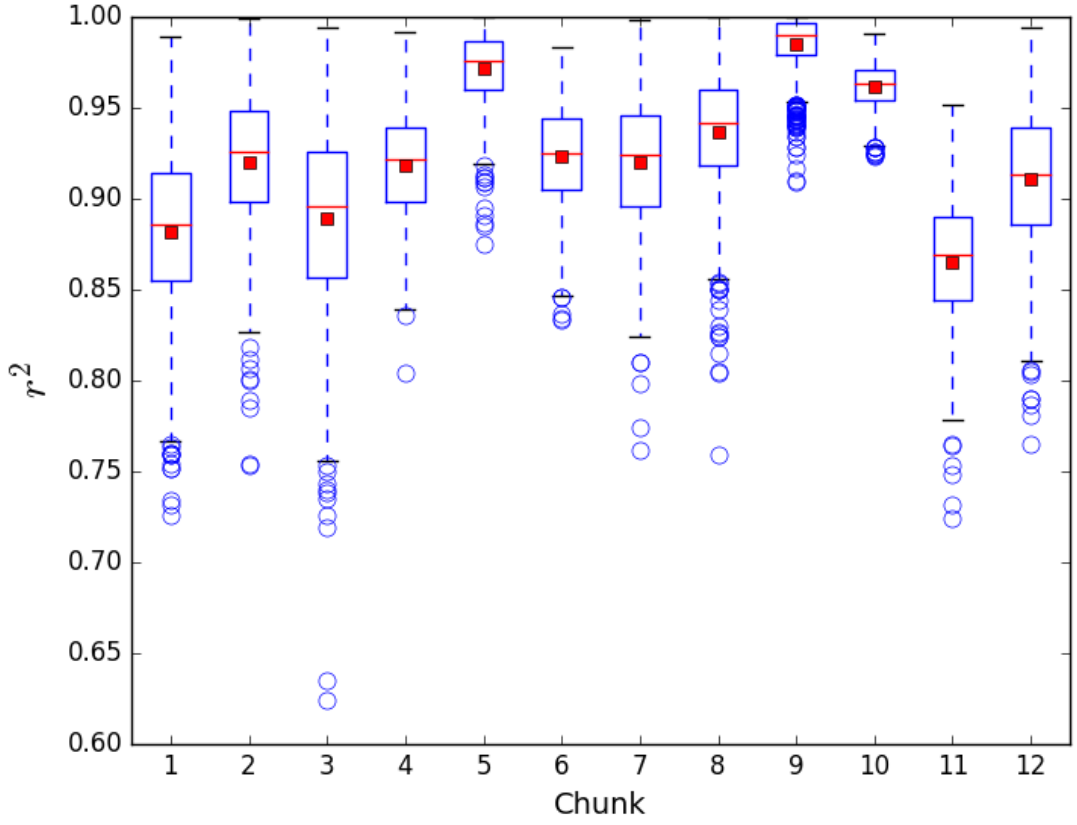
chunk#	interval	overall concordance (%)	number of SNPs both in the genotype file and reference table in the analysis region (type 2)	number of SNPs in phased reference haplotypes only (type 0)	number of SNPs in unphased study genotypes only (type 3)	imputation time
1	16000000-19000000	92.4	183	76075	4	12 h 2 min
2	19000000-22000000	93.9	144	76385	2	12 h 32 min
3	22000000-25000000	93.3	156	103121	7	15 h 3 min
4	25000000-28000000	94.4	332	98519	4	14 h 23 min
5	28000000-31000000	98.2	255	81171	13	5 h 33 min
6	31000000-34000000	96.5	254	88564	3	6 h 24 min
7	34000000-37000000	94.6	183	97716	0	7 h 5 min
8	37000000-40000000	95.7	257	95590	6	6 h 56 min
9	40000000-43000000	99.1	208	85321	7	6 h 9 min
10	43000000-46000000	97.6	1036	104230	27	7 44 min
11	46000000-49000000	94.0	326	112935	4	17 h 13 min
12	49000000-52000000	93.6	186	87070	0	6 h 11 min
whole chromosome 22	16000000-52000000	-	3520	1106697	77	17 h 13 min (chunks were calculated in parallel)

The table shows that the overall concordance was between 92.4 % and 99.1 % giving the average of 95.3 %. The average of 95.3 % is rather good as the recommendation is that the overall concordance should be around 95 %. The weakest overall concordance was 92.4 % and that is from the first chunk of the imputation. The imputation started from chromosome position 160 00 000 and the first input data SNP was at location 16 953 560, which left 953 560 nucleic strand positions before the first SNP to find its match. The first reference SNP that was imputed was at position 16 050 075.

We can see from the table that most of the input SNPs found their match from the reference data – which is beneficial in terms of imputation quality. The more overlapping shared SNPs there are between the input and reference data the more accurate the imputation should be.

IMPUTE2 defines the R^2 measurement as the squared correlation between input and masked/imputed genotypes at a SNP. The closer the value is to 1, the higher is the correlation between the input and masked/imputed genotype and that SNP has been imputed with high certainty. Figure 5 shows the squared correlation distribution of the individuals from imputed chunks. We can see from the figure that all the squared distributions were above 0.6. IMPUTE2 does not give a specific value for a cut-off above which the imputation results can be considered plausible, but states that cut-offs of 0.3 and 0.5 have been used by various groups.

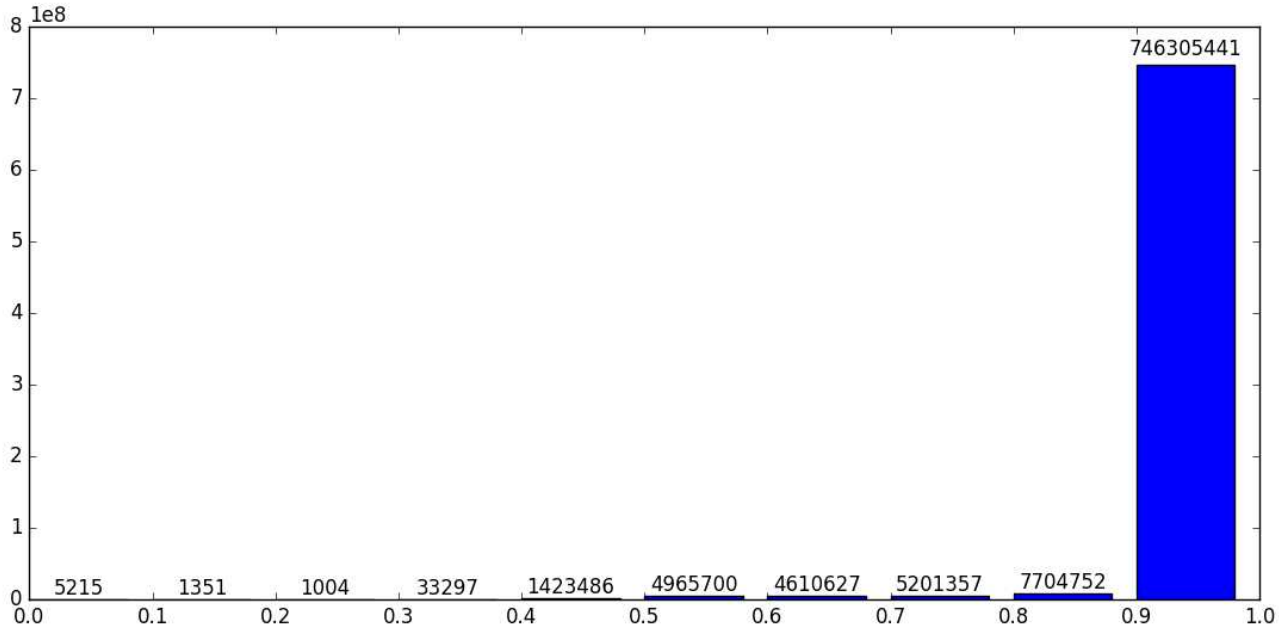
Figure 5. The squared correlation distribution of the individuals from imputed chunks shown as box plot. The plot shows that all $R^2 > 0.6$. The red box stands for mean and red line for median. (Calculated from r2_type0 from info_by_sample-file of each chunk)



5.2 Maximum values of imputed SNP probability triplets

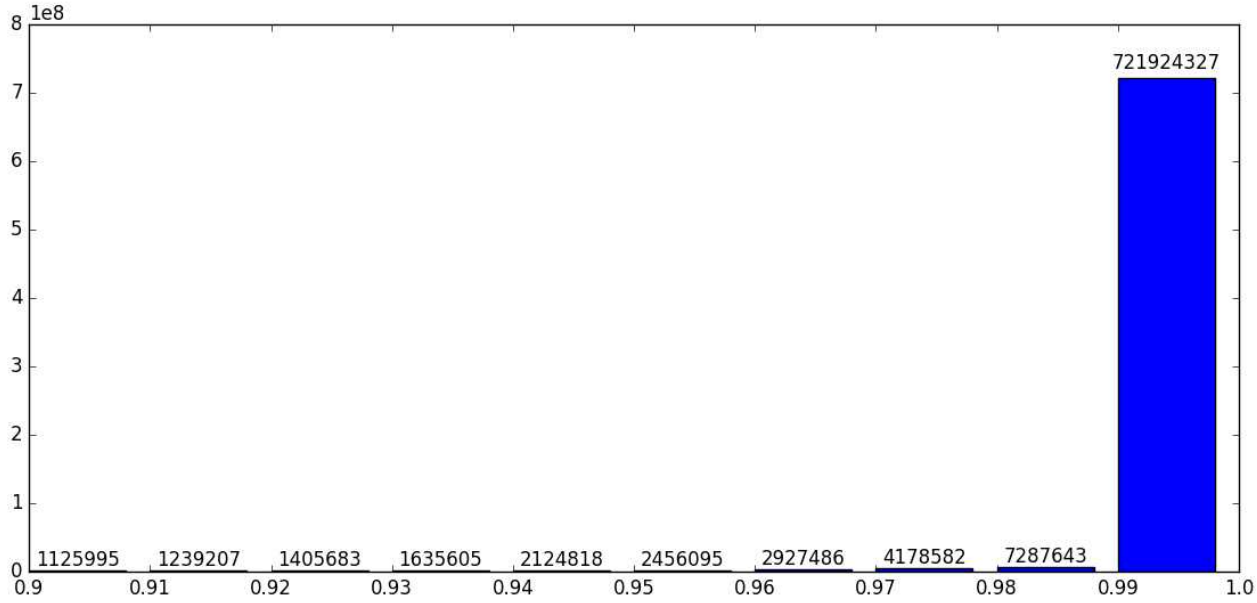
In total 746 305 441 of the 770 261 112 imputed genotype triplets of SNPs had the maximum probability between 0.9-1.00 for a genotype (Figure 6).

Figure 6. Frequencies of maximum probability estimates values of each triplet of all imputed SNPs to chromosome 22. Most of SNPs has the maximum probability of 0.9-1.00 for a genotype.



To further study the distribution of the maximum probabilities of the genotype triplets a histogram was plotted with probability values between 0.9-1.0 (Figure 7).

Figure 7. Frequencies of maximum probability estimates values of each triplet having a probability value between 0.9-1.0. Most of SNPs has the maximum probability of 0.99-1.00 for a genotype.



And, further studies showed that the median of the maximum probabilities for each triplet was 1.0. In total of 90 % of the imputed SNP genotype triplets had 1.0 as a probability for its most likely genotype.

Figures 6 and 7 show that majority of the SNPs that were imputed from the 1000G were given a genotype with a high probability. This shows that IMPUTE2 found haplotypes from the reference data to be matching with high confidence with the original genotyped SNPs.

5.3 Imputation times requirements

Table 5 also shows the imputation times. Each chunk took few hours to be imputed, the average imputation time for a chunk was 9 hours and 46 minutes, and the whole chromosome was imputed in 17 hours and 13 minutes. Naturally this is not an absolute, static, time requirement of the process. The imputation time depends on the environment and the settings of the imputation and of the possible other workloads from other tasks executing on the servers. However, the 17 hours and 13 minutes that this imputation took, does give an idea of how computationally laborious the imputation was in this case.

In this study, only chromosome 22 of KBCP was imputed. Chromosome 22 is the second shortest somatic chromosome. The imputation was performed between the interval of 16 000 000 - 52 000 000, because the first study SNP was located at 16 953 560. The interval of 0 - 15 999 999 was not imputed, as the interval did not contain any study SNPs. To perform imputation on the whole genome (female, XX) with the assumption that there would be enough study SNPs covering the whole length of the genome in chunks of 3 000 000 base pairs, there would be 1 011 chunks to be imputed (Table 6). If we take the average imputation time for the imputed chunks from chromosome 22 (9 h 46 min) and multiply that with the number of theoretical chunks for the whole genome, the expected time requirement would be (586 min * 1011) 13,7 months (592 446 minutes) without using parallelization.

Table 6. The lengths of chromosomes (female, XX) from Human Genome Assembly GRCh37[28] and the theoretical number of imputation chunks in relation to the chromosome size.

Chromosome	total length (bp)	theoretical number of chunks (size 3 million bp)
1	249 250 621	83
2	243 199 373	81
3	198 022 430	66
4	191 154 276	64
5	180 915 260	60
6	171 115 067	57
7	159 138 663	53
8	146 364 022	49
9	141 213 431	47
10	135 534 747	45
11	135 006 516	45
12	133 851 895	44
13	115 169 878	38
14	107 349 540	36
15	102 531 392	34
16	90 354 753	30
17	81 195 210	27
18	78 077 248	26
19	59 128 983	20
20	63 025 520	21
21	48 129 895	16
22	51 304 566	17
X	155 270 560	52
all	**3 036 303 846**	**1 011**

In this study, the imputation of chromosome 22 was performed by utilizing parallelization. Without parallelization, the imputation would theoretically have taken 4 days and 21 hours - that is 6.8 times longer than it took in practice. Thus, if we scale the theoretical imputation time of 1011 chunks in 13.7 months by dividing it with parallelization factor 6.8 the resulting time would decrease to 2 months.

In reality, the number of chunks would be smaller since the SNPs from iCOGS do not cover the whole genome area. As in the situation with chromosome 22, where the SNPs from iCOGS were all in an interval that covered only 67 % of the length of chromosome 22. Therefore, a

rough estimate of the imputation time of KBCP whole genome genotype data would be 1.5 months using the settings applied in this thesis.

Fortunately, imputation is highly parallelizable and by using more computational power, i.e. by having more CPU cores to utilize, the execution of imputation becomes faster. Theoretically the imputation can be completed in the imputation time of 1 chunk (if there were as many threads available as there are chunks). The parallelization factor of 6.8 resulted from the relatively limited capacity available at UEF Bioinformatics Center's cluster. However, The Sisu supercomputer of CSC offers 19200 CPU cores for a single task[29]. Therefore, the imputation can be parallelized down to time required to impute the single most time demanding chunk.

6 Discussion and Conclusion

In this study, chromosome 22 from KBCP -genotype data was imputed using the IMPUTE2 program and 1000G as reference. The experiment produced an imputed data set of the originally genotyped 3 597 SNPs extended with the imputation probabilities of 1 106 697 initially ungenotyped SNPs inferred from 1000G. Most of the SNPs that were imputed from the 1000G were given a genotype with a high probability (median 1.0).

IMPUTE2 provides quality measures utilizing the input data that can be used to evaluate the imputation quality. These quality measures are the overall concordance and squared correlation R^2.

The overall concordance was between 92.4 % and 99.1 % giving the average of 95.3 %. The average of 95.3 % can be considered rather good as the recommendation is that the overall concordance should be around 95 %. The individual level squared correlations of the imputed chunks were over 0.6, which is higher than what IMPUTE2 states as used cut-offs 0.3 and 0.5 which have been used by other groups.

The estimated imputation time for the whole KBCP iCOGS genotype genome data would be around 1.5 months with the UEF Bioinformatics Center's cluster. However, the imputation time can be substantially decreased with using more computational capacity, since imputation is a highly parallelizable process as the imputation is executed in chunks.

Performing an imputation run with IMPUTE2 is quite straightforward, especially when a single ready-made reference data set is used. The most tedious part was the pre-processing – transforming the data into genotype file format. And, naturally also figuring out the nature of the data itself that was to be imputed.

The pre-processing of the KBCP genotype data was executed with Python as the scripting language with data analysis libraries, e.g., pandas. Nowadays, Python is a very popular modern scripting tool used in many fields of scientific analysis. Previously I have had some experience with Python scripting, but for me pandas were a new library to work with. In addition, to learn to perform imputation with IMPUTE2, I also decided to use this thesis work as an opportunity to learn other techniques as well, e.g., pandas with Python.

Another technique that I learned was using a Slurm cluster and dividing tasks to Slurm batches. Studies performed on genetic datasets with large reference data tends to be computationally demanding. Therefore, high-performance server clusters are crucial and the knowledge to utilize them properly. Executing the data in batches at Slurm enabled parallelization of the runs, and faster generation of analysis results.

There was not much metadata connected to the KBPC genotype dataset. There was just the genotype file and the SNP location file. For example, the major and minor alleles of the genotypes were not in any order in the genotype file. Sometimes the major alleles were the first in the genotype pair and sometimes it was the latter. Even though this feature did not result in any problems in my knowledge, it still raised questions of what in the data might be easily misunderstood and therefore misused.

The way this study was conducted was in essence mainly reading the instructions and figuring things out quite independently. As bioinformatics is still quite a new field of science it is not taught in every University to a great extend. And even though there might be some bioinformatics taught, the definition of bioinformatics seems to vary a lot – covering anything from gene expression data analysis to model organisms. And, to my experience many of the bioinformaticians in laboratories are somewhat self-taught. Therefore, this kind of self-teaching approach to writing a thesis on a bioinformatics topic is very natural.

The schedule for this study was quite tight and studies like this could certainly be extended. I have now gone through the basic pipeline, but the actual outputs so far do not mean much at this point. Performing an association study with the imputed data would have been a natural next step. That would have probably shown whether the results would correlate with the already known susceptibility variants in the data. If not, that would have been a clear sign of the data not being credible.

Imputation accuracy depends on catching the right haplotypes for each SNP from the reference data. Naturally, to be able to find the right haplotype that haplotype needs to be present in the reference data. It is generally considered that the bigger the reference data set is the better. Of course, the bigger the reference data set is the more computationally demanding the imputation process becomes. The time of imputation becomes longer and the cost of imputation also increases (due to increased consumption of resources such as bought computation time and in the end electricity).

It is not clear whether imputation of the KBCP genotype data using only Finnish reference data would be a practical approach. It seems that when the Finnish variants are studied it is most important to have as much as possible of the Finnish individuals present in the reference data set and as large a reference data set as possible.

IMPUTE2 has not been updated since December 2014. This might be a sign that the program is not anymore under active development. For any possible future imputation plans, other imputation programs should be considered as well. For example, Beagle still seems to be under development, as the latest update on Beagle was just on the 8th of June 2017[30].

References

1. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).

2. Finnish Cancer Registry 2014 Finland. Available at: http://stats.cancerregistry.fi/stats/fin/vfin0004i0.html. (Accessed: 17th May 2017)

3. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100,** 57–70 (2000).

4. Cancer, I. A. for R. on. *WHO Classification of Tumours of the Breast*. (World Health Organization, 2012).

5. Aittomäki, K. *Lääketieteellinen genetiikka / toimituskunta: Kristiina Aittomäki, päätoimittaja, Jukka Moilanen, Markus Perola ; kirjoittajat: Aarnisalo Antti [ja 41 muuta]*. (Kustannus Oy Duodecim, 2016).

6. Human Genome Assembly GRCh38.p10 - Genome Reference Consortium. Available at: https://www.ncbi.nlm.nih.gov/grc/human/data. (Accessed: 20th May 2017)

7. Ezkurdia, I. *et al.* Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum. Mol. Genet.* **23,** 5866–5878 (2014).

8. Strachan, T. & Read, A. *Human Molecular Genetics, Fourth Edition*. (Garland Science, 2010).

9. Low Vision Resources Center — Genetics and Age-Related Macular Degeneration. Available at: http://lowvision.preventblindness.org/health-and-nutrition/genetics-and-age-related-macular-degeneration/. (Accessed: 4th June 2017)

10. Hoffmann, T. J. & Witte, J. S. Strategies for Imputing and Analyzing Rare Variants in Association Studies. *Trends Genet.* **31,** 556–563 (2015).

11. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genet.* **5,** e1000529 (2009).

12. IMPUTE2. Available at: https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#required_args. (Accessed: 19th April 2017)

13. Genotype imputation software tools | GWAS analysis - OMICtools. Available at: https://omictools.com/genotype-imputation-category. (Accessed: 4th June 2017)

14. Liu, Q. *et al.* Systematic assessment of imputation performance using the 1000 Genomes reference panels. *Brief. Bioinform.* **16,** 549–562 (2015).

15. 1000GP Phase 1 haplotypes 9 Dec 2013. Available at: https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html. (Accessed: 30th May 2017)

16. Ramnarine, S. *et al.* When Does Choice of Accuracy Measure Alter Imputation Accuracy Assessments? *PLoS ONE* **10,** (2015).

17. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47,** 373–380 (2015).

18. Day, F. R. *et al.* Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat. Genet.* **49,** 834–841 (2017).

19. iCOGS - Centre for Cancer Genetic Epidemiology. Available at: http://ccge.medschl.cam.ac.uk/research/consortia/icogs/. (Accessed: 15th April 2017)

20. Welcome to Python.org. Available at: https://www.python.org/. (Accessed: 28th May 2017)

21. McKinney, W. & others. Data structures for statistical computing in python. in *Proceedings of the 9th Python in Science Conference* **445,** 51–56 (van der Voort S, Millman J, 2010).

22. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **13,** 22–30 (2011).

23. gwas file format. Available at: http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html. (Accessed: 29th May 2017)

24. Kaviar Genomic Variant Database | SNP database | SNV database | ISB. Available at: http://db.systemsbiology.net/kaviar/. (Accessed: 29th May 2017)

25. Bioinformatics Center | UEF. Available at: http://www.uef.fi/en/web/bioinformatics. (Accessed: 30th May 2017)

26. Index of ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. Available at: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. (Accessed: 30th May 2017)

27. Slurm Workload Manager. Available at: https://slurm.schedmd.com/overview.html. (Accessed: 30th May 2017)

28. Human Genome Assembly GRCh37 - Genome Reference Consortium. Available at: https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh37. (Accessed: 18th June 2017)

29. CSC - 1.1 Sisu supercomputer. Available at: https://research.csc.fi/sisu-supercomputer. (Accessed: 18th June 2017)

30. Beagle 4.1. Available at: https://faculty.washington.edu/browning/beagle/beagle.html. (Accessed: 9th June 2017)

appendix 1: Example summary output of IMPUTE2

input genotype file: *KBCP_gff_chr22.csv*

input SNP location file: */KBCP_gff_chr22_strand_info.csv*

interval: *16000000 19000000*

==========

IMPUTE version 2.3.2

==========

Copyright 2008 Bryan Howie, Peter Donnelly, and Jonathan Marchini
Please see the LICENCE file included with this program for conditions of use.

The seed for the random number generator is 1781834084.

Command-line input: ./impute2 -m ./1000GP_Phase3/genetic_map_chr22_combined_b37.txt -h ./1000GP_Phase3/1000GP_Phase3_chr22.hap -l ./1000GP_Phase3/1000GP_Phase3_chr22.legend -g ./Imputation_test1/KBCP_gff_chr22.csv -strand_g ./Imputation_test1/KBCP_gff_chr22_strand_info.csv -int 16000000 19000000 -Ne 20000 -o /home/users/ipollane/imputointi/imputed_chr22/impute_bach_1.impute2

Nomenclature and data structure

Panel 0: phased reference haplotypes
Panel 2: unphased study genotypes

For optimal results, each successive panel (0,1,2) should contain a subset of the SNPs in the previous panel. When the data structure deviates from this ideal configuration, IMPUTE2 tries to use as much of the available information as possible; see documentation for details.

---------- Input files ----------

Panel 0 haplotypes: ./1000GP_Phase3/1000GP_Phase3_chr22.hap
Panel 0 hap legend: ./1000GP_Phase3/1000GP_Phase3_chr22.legend
Panel 2 genotypes: ./Imputation_test1/KBCP_gff_chr22.csv
Panel 2 strand info: ./Imputation_test1/KBCP_gff_chr22_strand_info.csv
genetic map: ./1000GP_Phase3/genetic_map_chr22_combined_b37.txt

---------- Output files ----------

main output: /home/users/ipollane/imputointi/imputed_chr22/impute_bach_1.impute2
SNP QC info: /home/users/ipollane/imputointi/imputed_chr22/impute_bach_1.impute2_info
sample QC info: /home/users/ipollane/imputointi/imputed_chr22/impute_bach_1.impute2_info_by_sample
run summary: /home/users/ipollane/imputointi/imputed_chr22/impute_bach_1.impute2_summary
warning log: /home/users/ipollane/imputointi/imputed_chr22/impute_bach_1.impute2_warnings

---------- Data processing ----------

-reading genetic map from -m file
--filename=[./1000GP_Phase3/genetic_map_chr22_combined_b37.txt]
--read 2498 SNPs in the analysis interval+buffer region

-reading strand info for Panel 2 from -strand_g file

--filename=[./Imputation_test1/KBCP_gff_chr22_strand_info.csv]
--read strand info for 197 SNPs in the analysis region

-reading Panel 2 genotypes from -g file
--filename=[./Imputation_test1/KBCP_gff_chr22.csv]
--detected 696 individuals
--read 197 SNPs in the analysis interval+buffer region

-using -strand_g file to align Panel 2 allele labels
--flipped strand at 108 out of 197 SNPs

-reading Panel 0 haplotypes from -h and -l files
--filename=[./1000GP_Phase3/1000GP_Phase3_chr22.hap]
--filename=[./1000GP_Phase3/1000GP_Phase3_chr22.legend]
--detected 5008 haplotypes
--read 84611 SNPs in the analysis interval+buffer region

-removing SNPs that violate the hierarchical data requirements
--no SNPs removed

-removing reference-only SNPs from buffer region
--removed 8343 SNPs

-checking strand alignment between Panel 2 and Panel 0 by allele labels
--flipped strand due to allele mismatch at 0 out of 197 SNPs in Panel 2

-aligning allele labels between panels

-removing non-aligned genotyped SNPs
--removed 0 out of 193 SNPs with data in multiple panels

Data summary

[type 0 = SNP in Panel 0 only]

[type 1 = SNP in Panel 1]

[type 2 = SNP in Panel 2 and all ref panels]

[type 3 = SNP in Panel 2 only]

-Upstream buffer region

--0 type 0 SNPs

--0 type 1 SNPs

--0 type 2 SNPs

--0 type 3 SNPs

--0 total SNPs

-Downstream buffer region

--0 type 0 SNPs

--0 type 1 SNPs

--10 type 2 SNPs

--0 type 3 SNPs

--10 total SNPs

-Analysis region (as defined by -int argument)

--76075 type 0 SNPs

--0 type 1 SNPs

--183 type 2 SNPs

--4 type 3 SNPs

--76262 total SNPs

-Output file

--76075 type 0 SNPs

--0 type 1 SNPs

--183 type 2 SNPs

--4 type 3 SNPs
--76262 total SNPs

-In total, 76272 SNPs will be used in the analysis, including 193 Panel 2 SNPs

-making initial haplotype guesses for Panel 2 by phasing hets at random and imputing missing genotypes from allele freqs

-setting storage space
-setting mutation matrices
-setting switch rates

Run parameters
---

reference haplotypes: 5008 [Panel 0]
study individuals: 696 [Panel 2]
sequence interval: [16000000,19000000]
buffer: 250 kb
Ne: 20000
input call thresh: 0.900
burn-in MCMC iterations: 10
total MCMC iterations: 30 (20 used for inference)
HMM states for phasing: 80 [Panel 2]
HMM states for imputation: 500 [Panel 0->2]

Run log

RESETTING PARAMETERS FOR "SURROGATE FAMILY" MODELING
-setting mutation matrices
-setting switch rates

diploid sampling success rate: 0.951

haploid sampling success rate: (no haploid sampling performed)

Imputation accuracy assessment

The table below is based on an internal cross-validation that is performed during each IMPUTE2 run. For this analysis, the program masks the genotypes of one variant at a time in the study data (Panel 2) and imputes the masked genotypes by using the remaining study and reference data. The imputed genotypes are then compared with the original genotypes to produce the concordance statistics shown in the table. You can learn more about this procedure and the contents of the table at http://mathgen.stats.ox.ac.uk/impute/concordance_table_description.html.

In the current analysis, IMPUTE2 masked, imputed, and evaluated 127237 genotypes that were called with high confidence (maximum probability >= 0.90) in the Panel 2 input file (-g or -known_haps_g).

When the masked study genotypes were imputed with reference data from Panel 0, the concordance between original and imputed genotypes was as follows:

Interval	#Genotypes	%Concordance	Interval	%Called	%Concordance
[0.0-0.1]	0	0.0	[>= 0.0]	100.0	92.4
[0.1-0.2]	0	0.0	[>= 0.1]	100.0	92.4
[0.2-0.3]	0	0.0	[>= 0.2]	100.0	92.4
[0.3-0.4]	68	35.3	[>= 0.3]	100.0	92.4
[0.4-0.5]	1350	46.8	[>= 0.4]	99.9	92.5
[0.5-0.6]	5538	54.6	[>= 0.5]	98.9	93.0
[0.6-0.7]	5619	66.7	[>= 0.6]	94.5	94.7
[0.7-0.8]	6514	76.5	[>= 0.7]	90.1	96.1
[0.8-0.9]	8653	85.2	[>= 0.8]	85.0	97.3
[0.9-1.0]	99495	98.3	[>= 0.9]	78.2	98.3