UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Technology

Anton Katsuba

Evaluating SNP Detection in Genomic Data: A Study on Mutation Patterns in the Estonian Biobank Data

Bachelor's Thesis (12 ECTS)

Curriculum Science and Technology

Supervisor:
PhD Vasili Pankratov

Tartu 2023

Abstract

Evaluating SNP Detection in Genomic Data: A Study on Mutation Patterns in the Estonian Biobank Data

Single nucleotide variants provide a rich source of information on cellular and genomic processes and phenomena. Rare variants are especially of interest, since they are likely to have experienced the least influence from evolutionary forces, which is important in estimating mutation rates. However, high quality SNV data is necessary for this. The goal of this thesis is to examine the quality of newly discovered SNVs in the Estonian Biobank dataset by looking at mutation patterns. The findings show that the subset of newly discovered SNVs has significant quality issues compared to the subset of already known SNVs. These issues are likely attributable to mapping errors.

Keywords: rare variants, mutation rate, whole genome sequencing, quality control

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Kokkuvõte

SNP tuvastamise hindamine genoomsetes andmetes: uuring mutatsioonimustrite kohta Eesti Biopanga andmetes

Üksiknukleotiidi variandid pakuvad rikkalikku teavet raku- ja genoomiprotsesside ning nähtuste kohta. Haruldased variandid on eriti huvipakkuvad, kuna neil on tõenäoliselt olnud vähem mõju evolutsioonilistele jõududele, mis on oluline mutatsioonikiiruste hindamisel. Kuid selleks on vajalikud kõrge kvaliteediga SNV andmed. Käesoleva uurimistöö eesmärk on uurida uues avastatud SNV-de kvaliteeti Eesti Biopanga andmekogumikus, analüüsides mutatsioonimustreid. Uurimustulemused näitavad, et uute avastatud SNV-de alamhulkadel on olulisi kvaliteediprobleeme võrreldes juba teadaolevate SNV-de alamhulkadega. Neid probleeme on tõenäoliselt põhjustanud kaardistamisvigad.

Keywords: haruldased variandid, mutatsioonikiirus, kogu genoomi sekveneerimine, kvaliteedikontroll

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Table of contents

Abstract 2

Kokkuvõte 3

Table of contents 4

Terms and abbreviations 5

1. Introduction 6

2. Literature review 8

    2.1. Identification of single nucleotide polymorphisms 8

    2.2. Sources of errors in SNP calling 9

    2.3. Estimating mutation rates 11

Aims of the thesis 13

3. Materials and methods 14

    3.1. Data description 14

    3.2. Methods 14

4. Results and discussion 16

    4.1. Mutation patterns and their differences 16

    4.2. Possible explanations 21

5. Summary 22

References 23

Non-exclusive license to reproduce thesis and make thesis public 25

Terms and abbreviations

AC — allele count

HTS — high-throughput sequencing

SNP — single nucleotide polymorphism

SNV — single nucleotide variant

NGS — next-generation sequencing

Ti/Tv — transitions to transversions ratio

WGS — whole genome sequencing

1. Introduction

The study of spontaneous mutations is essential in genetic research as these mutations can provide critical insights into human evolution and disease origins. One particular area of interest is rare mutations. These mutations can be especially informative as they play a crucial role in various genetic diseases and provide important markers for tracking population history. Despite the importance of having accurate estimates of mutation rates and their distribution across the genome, this area remains incompletely understood and subject to ongoing investigation.

One of the central resources that have proven invaluable for this purpose is biobanks, which store biological samples for use in research. The Estonian Biobank, one of the largest in Europe, provides a rich dataset that allows researchers to dive deep into the genetic landscape of a population. With its broad array of genetic data from thousands of individuals, the Biobank is uniquely poised to contribute significantly to our understanding of rare mutations.

The criticality of high-quality data cannot be understated in any research, and it takes on paramount importance in the field of genetic studies. When analyzing genetic data, the accuracy, precision, and reliability of the information gathered can significantly impact the findings and subsequent interpretations. The validity of results, the power to detect true genetic signals, and the success of subsequent interventions are all directly tied to the quality of the initial data.

Genetic studies often involve extensive datasets, which, if compromised in terms of quality, can lead to incorrect assumptions and findings. It is also worth noting that genetic data's inherent complexity necessitates strict quality control measures to ensure that the data is free of errors and biases. Anomalies, inaccuracies, and imprecisions in genetic data can significantly skew the results of analyses, leading to misleading conclusions and possibly faulty interventions.

Therefore, we decided to take a closer look at the Estonian Biobank whole genome sequencing data.

2. Literature review

2.1. Identification of single nucleotide polymorphisms

The advent of high-throughput DNA sequencing (HTS) has significantly expanded our capacity for genomic exploration. Prominent among its uses is the elucidation of single nucleotide polymorphisms (SNPs), which requires a sequence of operations, from raw genetic sequence extraction to data processing and utilization of a variety of tools.

The pipeline for SNP calling comprises a succession of tasks, from base calling and quality control to alignment/mapping and post-processing of the alignment. Subsequent stages entail quality score recalibration, variant and genotype calling, and filtering SNP candidates. It is essential to underscore that the selection of an alignment tool and its corresponding settings substantially influence the results. Inaccurately aligned reads can lead to artificial discrepancies from the reference, which can erroneously be labeled as SNPs in downstream processing. [1]

Likewise, it's standard practice to realign reads around small insertions and deletions (indels). Differences in addressing these indels could generate artificial SNPs in subsequent analyses. One of the primary concerns in SNP calling is the reduction of false-positive calls, and filtering stands as a crucial step in this regard. The typically employed filters examine for deviations from the Hardy–Weinberg equilibrium, min-max read depth, proximity to indels, strand bias, etc. These filters might also discard real SNPs from the candidate list, but they play a pivotal role in minimizing SNP calling artifacts. [1]

The translation of next-generation sequencing (NGS) image files into a set of identified SNPs encompasses several steps. These include image analysis, alignment and assembly, and eventually SNP and genotype calling. Genotype probabilities for a single individual can be calculated from alignments using recalibrated quality scores. [2]

Moreover, calling of SNPs and genotypes is optimally achieved using data from multiple individuals simultaneously, with the pattern of linkage disequilibrium employed for SNP and genotype calling whenever feasible. The analysis of low coverage data can proceed

by considering uncertainty in the genotype calls, rather than making assumptions about the correctness of any particular genotype call. The methodologies used for calling SNPs and for incorporating uncertainty in SNP genotypes can significantly impact downstream analyses, including association mapping analyses. [2]

SNPs serve as molecular markers in many studies involving monogenic or complex diseases due to their high frequency and binary variation pattern. For instance, the commonly used oral anticoagulant warfarin often shows high individual dose requirement variability leading to adverse effects. In a study, two polymorphisms in CYP2C9 were investigated in patients during long-term warfarin therapy, with the polymorphisms associated with an increased risk of over-anticoagulation and bleeding events. This exemplifies the potential biomedical applications of SNP identification and genotyping. [3]

Taken together, the identification and genotyping of SNPs constitute a multi-step process that relies heavily on the accuracy of high-throughput sequencing technologies. The insights gleaned from such efforts can significantly impact various biomedical fields, highlighting the crucial role of SNP analysis in our understanding of genetic variations in health and disease.

2.2. Sources of errors in SNP calling

Next-generation sequencing data can suffer from high error rates due to multiple factors, such as base-calling and alignment errors [2]. In low-coverage sequencing, for example, there is a high probability of sampling only one of the two chromosomes of a diploid individual at a given site. This creates difficulties for accurate SNP and genotype calling, leading to significant uncertainty associated with the results. Accounting for this uncertainty is essential as it influences downstream analyses, like identification of rare mutations, estimation of allele frequencies, and association mapping [2].

Different sequencing platforms have been associated with specific errors. For instance, the trimer GGT is the most error-associated trimer and shows variability across instruments. [4] GGT is far more over-represented in HiSeq 2500 errors than in any other platform, while

the NovaSeq 6000 platform seems less influenced by these motifs. In contrast, HiSeq 2500 and MiSeq exhibit more motif-dependent errors. [4]

Another common error mode in Illumina platforms is associated with homopolymers, sequences with runs of the same base. Illumina reads often substitute the first base after a homopolymer with the homopolymer base, which is linked to phasing issues. This error accounts for between 0.7 and 5.3% of all errors, depending on the base and platform [4].

Systematic errors in high-throughput sequencing data often show strong tendencies for specific base substitutions. For example, a marked propensity for T > G errors compared to all others has been observed. This reveals a higher substitution rate to G than other nucleotides, with the substitution rate to A or T considerably lower than the rate to C [5].

Errors can also arise due to variation in read mapping. When accounting for very rare events, mapping errors, rather than sequencing errors, contributed most to the artifacts. [6] The data highlighted a small number of differences between libraries made from the same DNA at almost the same time. Remapping reads around these sites with different software showed that most differences were due to variation in read mapping rather than inherent differences between libraries. [6]

The read preprocessing step has been shown not to improve the accuracy of variant calling, contrary to expectation. While trimming off low-quality tails from reads can help align more reads, it can also introduce false positives. [7] Moreover, the relative performance of three popular multi-sample SNP callers, SAMtools, GATK, and GlfMultiples, varied with sequencing depth. [7]

Monitoring the transition/transversion (Ti/Tv) ratio helps to assess the quality of SNP calls. A higher Ti/Tv ratio generally indicates higher accuracy. When detected variants demonstrate a ratio closer to the expected ratio for random substitutions (e.g., 0.5), it implies low-quality variant calling or data. [7]

2.3. Estimating mutation rates

Understanding mutation rates and their dependence on the surrounding nucleotide context is crucial in the fields of human genetics and evolutionary biology. A detailed grasp of these parameters not only informs our understanding of the genesis of genetic diseases but also aids in timing evolutionary events.

The process of estimating de novo mutation rates, particularly in the context of single nucleotide variants, is a complex task that requires high depth sequencing and trio information. De novo mutations refer to new mutations in an individual that are absent in their parents. The estimation of de novo mutation rates is essential because of the fundamental role these mutations play in causing genetic diseases and contributing to human evolution [8]. Nonetheless, given the rarity of de novo events, determining a precise estimate is challenging, and uncertainties remain regarding the actual rate and the influence of factors such as paternal age.

Estimating de novo insertion-deletion (indel) mutation rates is an even greater challenge due to their comparative rarity and the increased difficulty in their detection. To date, direct measurements of indel de novo mutation rates have been scant, with the only available data being from a study of a single trio [8].

A thorough understanding of genome-wide single-nucleotide germline mutation rates is crucial to studying human genome evolution. These rates are influenced by a variety of genomic features surrounding the mutation site, including replication timing, histone modifications, and recombination rate. Some of these factors suggest specific mutagenic mechanisms at play [9]. Interestingly, features like GC content, DNase hypersensitivity, CpG islands, and H3K36 trimethylation have been linked to both increased and decreased mutation rates, contingent on the nucleotide context. [9]

Family-based whole-genome sequencing (WGS) data is considered the gold standard for studying the human germline mutation rate, and has allowed researchers to accurately estimate the genome-wide average mutation rate. Nevertheless, the low inherent germline mutation rate makes it challenging to accumulate a sufficiently large dataset for fine-scale

estimation of mutation rates and the identification of factors explaining genome-wide mutation rate variability. [9]

The nucleotide context—the nucleotides flanking a polymorphic site—significantly impacts nucleotide substitution probabilities [10]. A study found that the inclusion of local nucleotides in a sequence context model improved its fit to the observed data. For example, a trinucleotide (3-mer) model with a single 5′ and 3′ nucleotide flanking the polymorphic position improved the fit to data over a 1-mer model. Further, the heptanucleotide (7-mer) model with three flanking nucleotides on each side performed better than both the 3-mer and the pentanucleotide (5-mer) models. [10]

Understanding mutation rates has important implications for studying human diseases and population history. For instance, mutation rate estimates are a crucial aspect in research seeking to establish the causality of sequence variants in human disease [11], and also in studies aiming to infer human population history from individual whole-genome sequences. [12]

Aims of the thesis

For this thesis, we have defined the following goals:

1. Find out whether there are unexpected substitution patterns within rare single nucleotide variants discovered in the Estonian Biobank whole genome sequencing dataset.
2. If there are, examine possible reasons behind them.

3. Materials and methods

3.1. Data description

In this study we used data derived from whole-genome sequences of 2695 Estonian Biobank participants [13, 14], sequenced using the Illumina technology and called against human reference genome version GRCh38. We started from a file tabulating all single nucleotide variant positions together with the reference and the alternative allele, alternative allele frequency in this dataset and an rsID annotation based on dbSNP version 155. Individual genotypes were not available and hence working with this data did not require ethics approval.

3.2. Methods

This set of SNVs was filtered (bcftools 1.16) based on alternative allele count (AC). Positions with AC = 1 were kept. Henceforth the SNVs with AC = 1 will be referred to as singletons. These singletons are expected to be a mixture of miscalled variants and very rare (and hence young) mutations. The latter condition ensures that this set of singletons is affected by selection as little as possible.

Human reference genome version GRCh38 was used to extract the information about the flanking nucleotides for each SNV position allowing us to identify the original 3-mer. We used the alternative allele as a proxy for the derived nucleotide to identify the nucleotide resulting from the mutation. Only autosomes were used for the analysis.

12.76 million singletons in a sample of approximately 2500 individuals were analyzed. Frequencies of occurrence of all triplets in the reference genome were calculated and compared to the frequencies of occurence of those triplets as mutation sites in the dataset of singletons. After that the dataset was split into two subsets: singletons with an assigned rsID (N = 12493267) and those without it (N = 262416). For each subset a substitution table was created: for every triplet the number of times it was substituted into a particular nucleotide was counted.

Chi-square test was performed on a 96x2 contingency table using the scipy (v1.10.1) Python package. This table was obtained from the above substitution count tables by discarding the cells where no substitution occurs and coalescing the triplets which are reverse complements of each other into the same category.

Moreover, the ratio of transitions to transversions was calculated as an additional quality metric.

4. Results and discussion

4.1. Mutation patterns and their differences

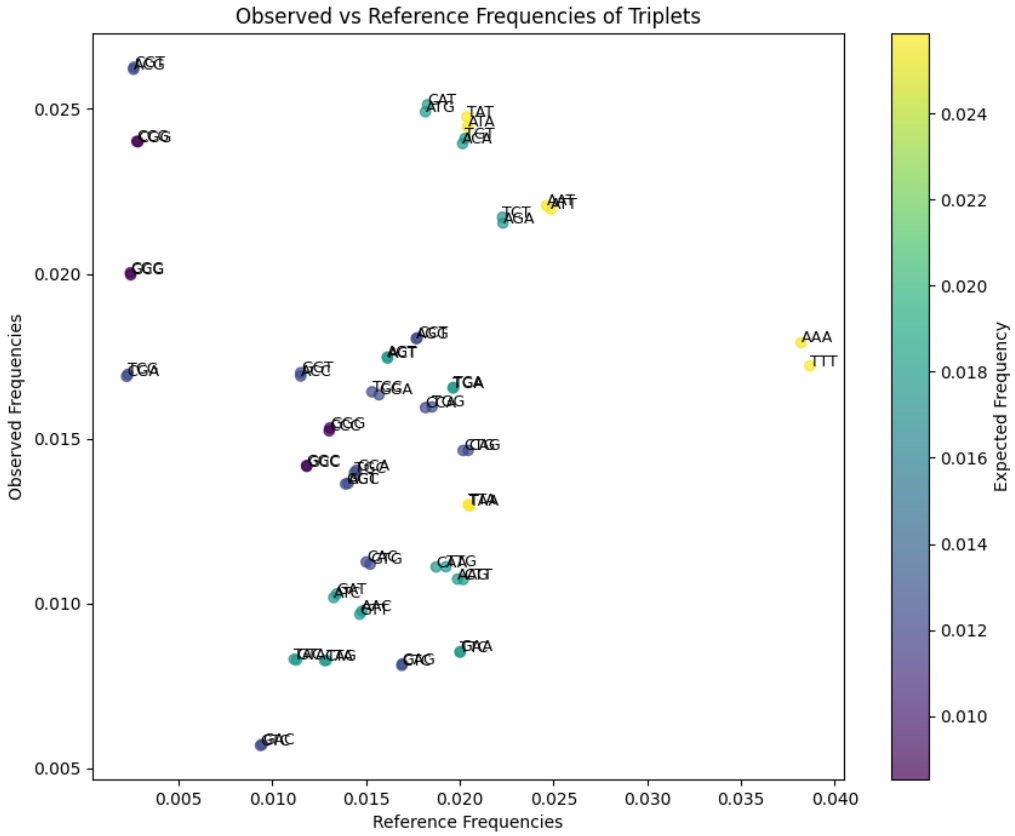

Figure 1: Comparing triplet frequencies. Color represents expected frequencies if triplet distribution were dictated by nucleotide frequencies.

Figure 1 represents the results of an initial sanity check necessary to ensure that all downstream analysis is correct. The graph shows the frequency of occurence of different triplets (3-mers) in the reference genome (x-axis) versus the same frequency among the subset of triplets carrying a singleton in the middle position (y-axis). The first notable thing is that reverse complements go together despite the analysis being done only on one strand of the reference genome. This is due to the so-called Chargaff's second parity rule. [15] The first rule which in effect describes nucleotide complementarity is widely known, however, its

analogue is true for single strands for a large part of known genomes, including the human one.

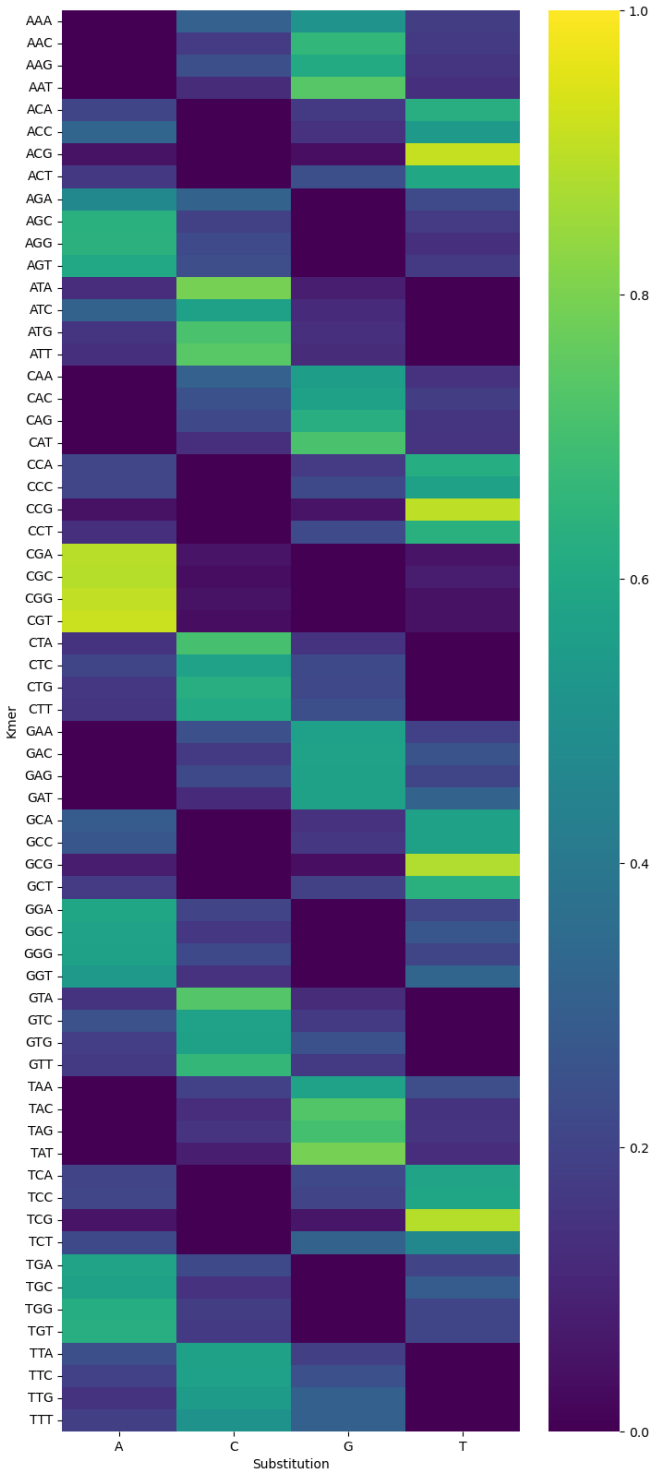

Figure 2: Substitution patterns in the rsID subset. Values are frequencies that sum up to one in each row.

The next notable thing is the excess of CGN/NCG triplets among singletons. CpG nucleotide combination is known to have a very high mutation rate, and seeing this shown on the graph is also in accordance with expectation.

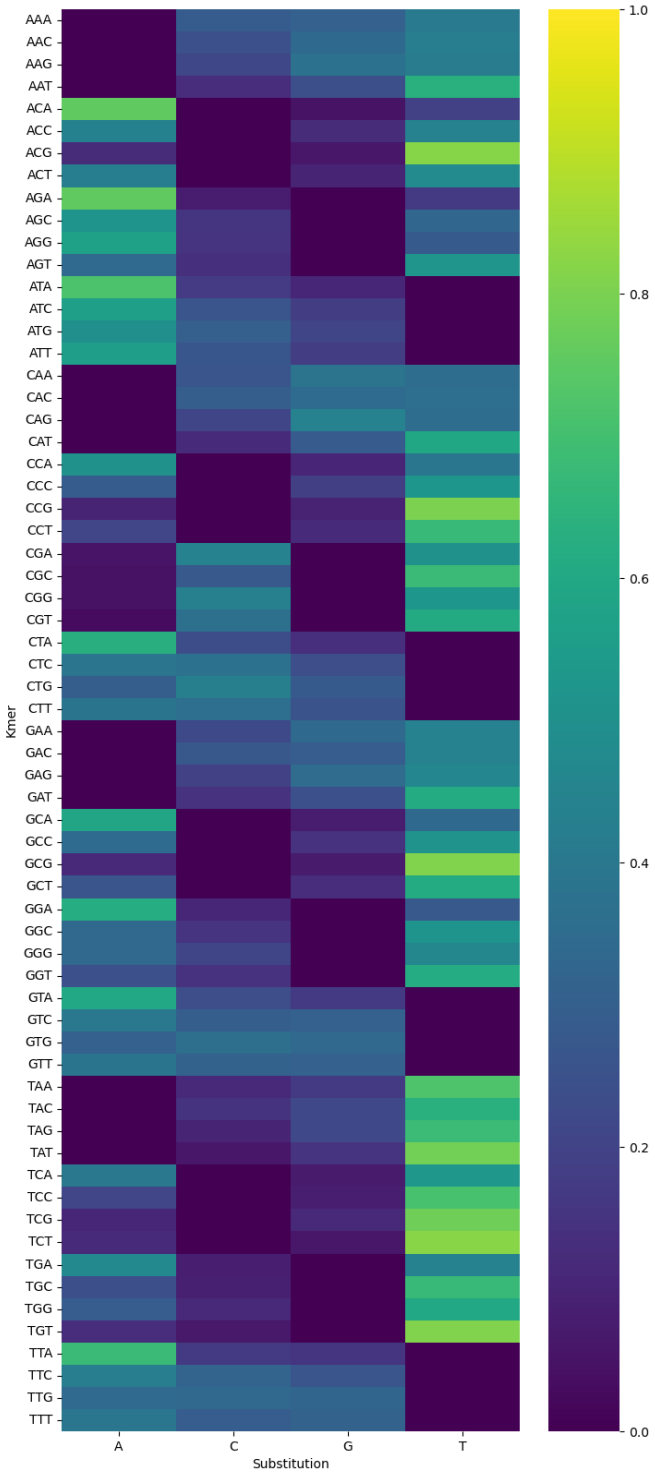

Figure 3: Substitution patterns in the non-rsID subset. Values are frequencies that sum up to one in each row.

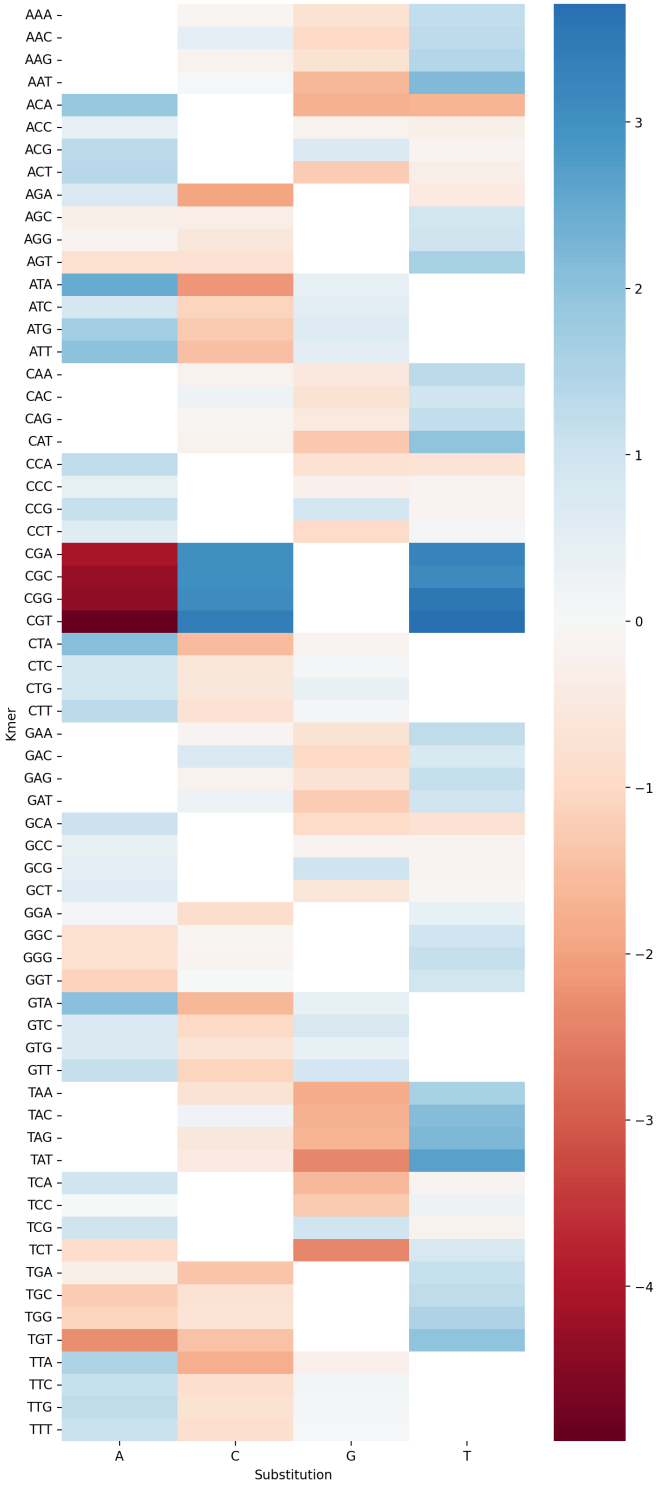

Figure 4: A comparison of substitution differences between the two SNVs subsets. The change is represented as log2 ratio of non-rsID to rsID frequencies.

Figure 2, Figure 3 and Figure 4 deal with substitution patterns across the subset of known and newly discovered SNVs. THe frequency of substitutions is calculated for each triplet across columns. In other words, the values sum up to unity in each row. The first two figures describe the patterns inherent to each subset, whereas the third figure compares them.

Figure 2 exhibits a normal mutational profile: there is a clear spike of G>A/C>T substitutions at CGN/NCG triplets, and most mutations are transitions.

Figure 3 exhibits a non-standard mutational profile: there is a puzzling lack of CGN/NCG involvement, and overall mutations are much more dispersed, there are significantly less transitions in comparison with the previous figure.

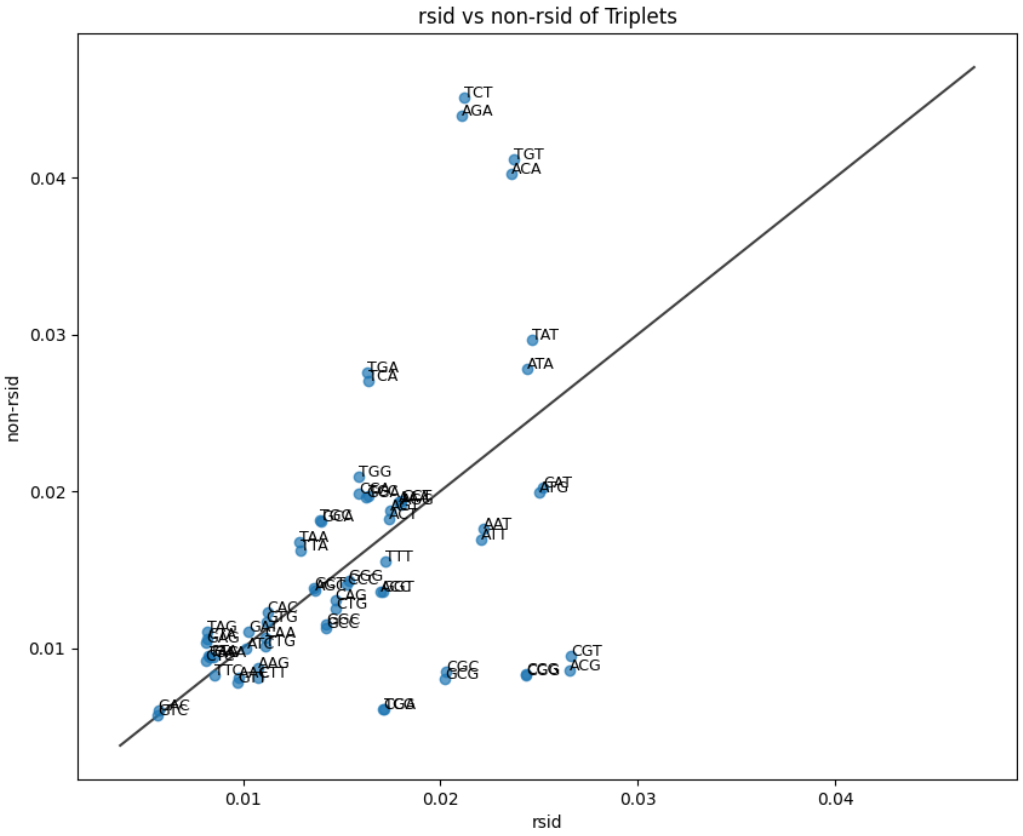

Figure 5: Comparing triplet frequencies between the rsID and non-rsID subsets analogously to Figure 1.

If we compute the transitions to transversions ratio for each subset (Ti/Tv), we see that the rsID subset has the value of Ti/Tv = 1.99, whereas the non-rsID subset has the value of Ti/Tv = 0.63. The values close to 2 are generally considered acceptable, meanwhile values close to 0.5 indicate a serious problem, since in this case the substitutions are nearly random. [7]

Performing the chi-squared test on the 96x2 contingency table described in methods gives a p-value $< .001$, which allows us to reject the null hypothesis of independence.

If we compare the frequencies of triplet occurence in the rsID and non-rsID subsets, we notice that the non-rsID subset mainly differs by the enrichment with the TCT, TGT, TGA, TAT triplets and their complements.

4.2. Possible explanations

What could be the cause of the observed quality issues in the non-rsID subset? Let's consider sequencing errors as the main driver. Unfortunately, the literature suggests that this is not the case. Sequencing error profile typical for Illumina systems does not involve ubiquitous substitutions to T, which is observed in our case. It does involve substitutions to G, which are not observed in high quantities. However, the most common erroneous substitution is T > G, and we can see on the Figure 4 that nearly all cases of possible T > G changes are colored light blue, which indicates 1.5x to 2x change in frequency. [5] This suggests some contribution of direct sequencing errors to the mutational profile of non-rsID singleton subset.

We are still left with a strange anomaly concerning the most mutable CpG-containing triplets. Despite the reduced frequency of G > A substitutions, we do not observe a corresponding decrease in frequency of C > T substitutions in reverse complements of said triplets. One can notice that the values in the Figure 2 and Figure 3 mirror each other horizontally when looking at reverse complementary triplets. This pattern breaks for the family of CGN triplets in the non-rsID table. This observation suggests some sort of strand asymmetry. Sometimes, there are reads mostly for one strand and not the other, this is called strand bias. Other times, there is asymmetrical G > T mutation due to the chemical environment during library preparation. However, it is difficult to estimate the possible contribution of these processes from the given data.

Third option is issues with read mapping. Some authors claim [6] that most issues with rare variants come specifically from mismapping.

5. Summary

Our study employed the dataset from the Estonian Biobank to investigate the quality of newly discovered single nucleotide variants. Using the human reference genome GRCh38, we examined 12.76 million singletons from 2695 individuals, and separated them into two groups: those with an assigned rsID and those without.

When examining the substitution patterns across both subsets, we observed an expected mutational profile in the rsID subset, but a non-standard profile in the non-rsID subset. There was a lack of CGN/NCG involvement and a significant decrease in the transitions to transversions (Ti/Tv) ratio compared to the rsID subset.

Although we considered sequencing errors as a potential driver of these observations, the data did not fully align with this hypothesis. Instead, our results suggest the presence of strand asymmetry, possibly due to strand bias or asymmetrical G > T mutation during library preparation. Read mapping issues may also contribute to these anomalies. Further research is required to understand the mechanisms behind these findings and their implications for obtaining quality data.

It it worth noting that the non-rsID subset comprises about 2% of all SNVs. Thus, the discovered issues do not constitute a serious problem for usability of biobank WGS data.

References

[1] A. Altmann, P. Weber, et al., "A beginners guide to SNP calling from high-throughput DNA-sequencing data," *Human Genetics*, vol. 131, no. 10, pp. 1541–1554, Oct. 1, 2012, doi: 10.1007/s00439-012-1213-z.

[2] R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song, "Genotype and SNP calling from next-generation sequencing data," *Nature Reviews Genetics*, vol. 12, no. 6, pp. 443–451, Jun. 2011, doi: 10.1038/nrg2986. [Online]. Available: https://www.nature.com/articles/nrg2986

[3] S. Kim, and A. Misra, "SNP genotyping: technologies and biomedical applications," *Annu. Rev. Biomed. Eng.*, vol. 9, no. 1, pp. 289–320, 2007, doi: 10.1146/annurev.bioeng.9.060906.152037. [Online]. Available: https://doi.org/10.1146/annurev.bioeng.9.060906.152037

[4] N. Stoler, and A. Nekrutenko, "Sequencing error profiles of illumina sequencing instruments," *NAR Genomics Bioinf.*, vol. 3, no. 1, Mar. 1, 2021, doi: 10.1093/nargab/lqab019. [Online]. Available: https://doi.org/10.1093/nargab/lqab019

[5] F. Meacham, D. Boffelli, et al., "Identification and correction of systematic error in high-throughput sequence data," *BMC Bioinf.*, vol. 12, no. 1, p. 451, Nov. 21, 2011, doi: 10.1186/1471-2105-12-451. [Online]. Available: https://doi.org/10.1186/1471-2105-12-451

[6] H. Li, "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data," *Bioinf.*, vol. 27, no. 21, pp. 2987–2993, Nov. 1, 2011, doi: 10.1093/bioinformatics/btr509. [Online]. Available: https://doi.org/10.1093/bioinformatics/btr509

[7] Q. Liu, Y. Guo, et al., "Steps to ensure accuracy in genotype and SNP calling from illumina sequencing data," *BMC Genomics*, vol. 13, no. 8, Dec. 17, 2012, doi: 10.1186/1471-2164-13-S8-S8. [Online]. Available: https://doi.org/10.1186/1471-2164-13-S8-S8

[8] S. Besenbacher, S. Liu, et al., "Novel variation and de novo mutation rates in population-wide de novo assembled danish trios," *Nature Commun.*, vol. 6, no. 1, p. 5969, Jan.

19, 2015, doi: 10.1038/ncomms6969. [Online]. Available: https://www.nature.com/articles/ncomms6969

[9] J. Carlson, A. E. Locke, et al., "Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans," *Nature Commun.*, vol. 9, no. 1, p. 3753, Sep. 14, 2018, doi: 10.1038/s41467-018-05936-5. [Online]. Available: https://www.nature.com/articles/s41467-018-05936-5

[10] V. Aggarwala, and B. F. Voight, "An expanded sequence context model broadly explains variability in polymorphism levels across the human genome," *Nature Genetics*, vol. 48, no. 4, pp. 349–355, Apr. 2016, doi: 10.1038/ng.3511. [Online]. Available: https://www.nature.com/articles/ng.3511

[11] D. G. MacArthur, T. A. Manolio, et al., "Guidelines for investigating causality of sequence variants in human disease," *Nature*, vol. 508, no. 7497, pp. 469–476, Apr. 2014, doi: 10.1038/nature13127. [Online]. Available: https://www.nature.com/articles/nature13127

[12] H. Li, and R. Durbin, "Inference of human population history from individual whole-genome sequences," *Nature*, vol. 475, no. 7357, pp. 493–496, Jul. 2011, doi: 10.1038/nature10231. [Online]. Available: https://www.nature.com/articles/nature10231

[13] V. Pankratov, F. Montinaro, et al., "Differences in local population history at the finest level: the case of the estonian population," *Eur. J. Human Genetics*, vol. 28, no. 11, pp. 1580–1591, Nov. 2020, doi: 10.1038/s41431-020-0699-4. [Online]. Available: https://www.nature.com/articles/s41431-020-0699-4

[14] M. Kals, T. Nikopensius, et al., "Advantages of genotype imputation with ethnically matched reference panel for rare variant association analyses," bioRxiv, 2019. [Online]. Available: https://www.biorxiv.org/content/10.1101/579201v2

[15] P. Fariselli, C. Taccioli, L. Pagani, and A. Maritan, "DNA sequence symmetries from randomness: the origin of the chargaff's second parity rule," *Briefings Bioinf.*, vol. 22, no. 2, pp. 2172–2181, Apr. 8, 2020, doi: 10.1093/bib/bbaa041. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7986665/

Non-exclusive license to reproduce thesis and make thesis public

I, Anton Katsuba,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) toreproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,
   Evaluating SNP Detection in Genomic Data: A Study on Mutation Patterns in the Estonian Biobank Data,
   supervised by Vasili Pankratov.
2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.
3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.
4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Anton Katsuba

24/05/2023