JOURNAL OF
EMERGING INVESTIGATORS

Article

The impact of genetic analysis on the early detection of colorectal cancer

Nikita Agrawal[1], Esmael Jafari Haddadian[2]
[1]Whitney M. Young Magnet High School, Chicago, Illinois
[2]University of Chicago, Biological Sciences Collegiate Division, Chicago, Illinois

SUMMARY

Although the 5-year survival rate for colorectal cancer is below 10%, it increases to greater than 90% if it is diagnosed early. We hypothesized from our research that analyzing non-synonymous single nucleotide variants (SNVs) in a patient's exome sequence would be an indicator for high genetic risk of developing colorectal cancer. First, the patient's exome sequence and the reference exome sequence were repeatedly aligned to identify the regions of similarity. The alignment between the two sequences that resulted in the most similar regions (optimal alignment) was selected. Next, we performed variant calling and identified variants in the patient's exome sequence. We applied a quality control check to assess sequencing data quality and filtered out the variants that did not pass the quality control check. Finally, the remaining selected variants were annotated with biologically pertinent information and explored for their potential roles in human disease by cross-referencing databases. A variant in the FGFR4 gene, known to cause accelerated cancer progression and tumor cell motility, was found in the patient's exome sequence. Studies suggest that this variant corresponds with an increased risk of colorectal cancer, supporting the usefulness of this procedure in early detection of colorectal cancer. This research was expanded to demonstrate that exome sequencing methods are capable of identifying other genetic variants. With higher computational power, one can produce more accurate alignment readings, detecting even the smallest of deviations from the reference exome sequence and thus enhancing the ability to evaluate genetic risk of disease.

INTRODUCTION

Early diagnosis of colorectal cancer improves survival. The 5-year survival rate for colorectal cancer that has not spread is 90% (1). The 5-year survival rate decreases to 72% if the cancer has spread to surrounding tissues, organs and/or the regional lymph nodes (2). Furthermore, if the cancer has spread to distant parts of the body, the 5-year survival rate is 14% (2). Screening is therefore imperative to detect colorectal cancer and to increase the 5-year survival rate for a patient.

In 2018, researchers cited colorectal cancer as the third most deadly and fourth most diagnosed cancer in the world (3). However, in 2020, colorectal cancer became the second most deadly and third most commonly diagnosed cancer in the world (4). This shows that colorectal cancer is becoming an increasingly serious threat. Although the rate of people diagnosed with colorectal cancer has declined overall since the mid-1980s (2), according to the Colon Cancer Coalition, the diagnosis rate for people under the age of 50 has increased alarmingly. An estimated 49 people under the age of 50 were diagnosed per day in 2020 with the disease (5). Despite this, most countries recommend colorectal cancer screening for individuals over the age of 50 (6). Early detection of colorectal cancer refers to screening: the process of looking for colorectal cancer in people who have no symptoms of the disease (7). Colorectal screening procedures aim to find polyps, or growths that appear on the surface of the colon (8, 9). These polyps may travel from the colorectal area to other parts of the body, such as the liver, in a phenomenon known as spreading. By finding polyps early, scientists hope to reduce mortality from colorectal cancer.

Currently, colonoscopy is the most commonly used colorectal cancer screening test. Other colorectal screening tests include sigmoidoscopy and virtual colonoscopy. It is recommended that individuals interested in colorectal cancer screening receive either a colonoscopy every 10 years, a sigmoidoscopy every 5 years, or a virtual colonoscopy every 5 years (7). According to the American Cancer Society, these tests are viable options for many patients particularly due to their affordability and safety (10). Colonoscopy and sigmoidoscopy both use a thin, flexible tube with a camera at the end to look at the colon. However, colonoscopy examines the entire colon, while sigmoidoscopy examines only the lower part of the colon (11). On the other hand, a virtual colonoscopy uses a CT scan to produce hundreds of cross-sectional images of the abdominal organs. The images are then combined and digitally manipulated to provide a detailed view of the inside of the colon and rectum (12).

A novel approach of detecting colorectal cancer currently under development uses genetic analysis (13). Although two people are 99.9% genetically identical when it comes to regular organ functions, the other 0.1% can explain why one person is more vulnerable to certain diseases than the other person (14). Genetic analysis entails analyzing a sample of DNA to look for mutations that may increase the risk of spreading the disease. The analysis compares the patient's DNA sequence to a healthy reference DNA sequence which does not have a high genetic risk for the disease. A DNA sequence consists of a linear string of nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G) (15). The differences in nucleotide bases between the patient's DNA sequence and the reference DNA sequence are known as single nucleotide variants (SNVs). Every three nucleotides in the DNA sequence code for an amino acid and the combination of these amino acids code for

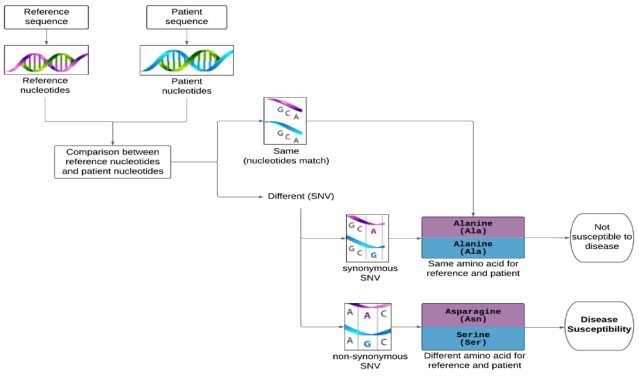

Figure 1: Flowchart of genetic analysis process. This figure depicts the logical sequence during genetic analysis leading to identification of a disease. The purple represents the reference sequence and the blue represents the patient sequence.

a protein. Although a change of nucleotide base may occur in a DNA sequence, it does not necessarily imply that the amino acid sequence will also change (16). For example, the two nucleotide sequences C-C-C and C-C-A both code for the amino acid proline. The mutations that do not alter the amino acid sequence are called synonymous SNVs. On the other hand, non-synonymous SNVs alter the amino acid sequence and make up the majority of known genetic diseases (**Figure 1**) (17).

Proteins are essential for the structure, function, and regulation of the body's tissues and organs (18). The patients may inherit a genetic change that increases their risk of colorectal cancer (germline mutation) or the patient may acquire the genetic change over their lifetime (somatic mutation). In such cases, screening using genetic analysis can help identify the change of nucleotide bases between the patient's DNA sequence and the reference DNA sequence and thus making it easier to pinpoint the source of colorectal cancer.

Although a human DNA sequence is exceptionally large, only around 1.5% encodes exons – segments of DNA that code for proteins (19). An exome sequence is a sequence of all of the exons in a DNA sequence (20). Analyzing the exome sequence can help determine if there are any colorectal cancer-related mutations in the proteins. These mutations within the exome sequence lead to the development of colorectal cancer (21).

The combined effect of genetic syndromes and family history may explain up to 30% of colorectal cancer susceptibility, whereas the remaining genetic risk of colorectal cancer may be accounted for by a combination of high-prevalence and low-penetrance of common genetic variants (22). Although a recent study performed using exome sequencing proposes genes contributing to a higher genetic risk for colorectal cancer, much needs to be explored in this area (23). Another study has used the genetic analysis approach for diagnosing patients with breast cancer (24).

Since a significantly smaller amount of DNA is sequenced when analyzing only the exons, exome sequencing is computationally inexpensive compared to sequencing an entire genome. Exome sequencing is a widely used Next Generation Sequencing method that is able to sequence DNA at unprecedented speeds (25). Our project is based

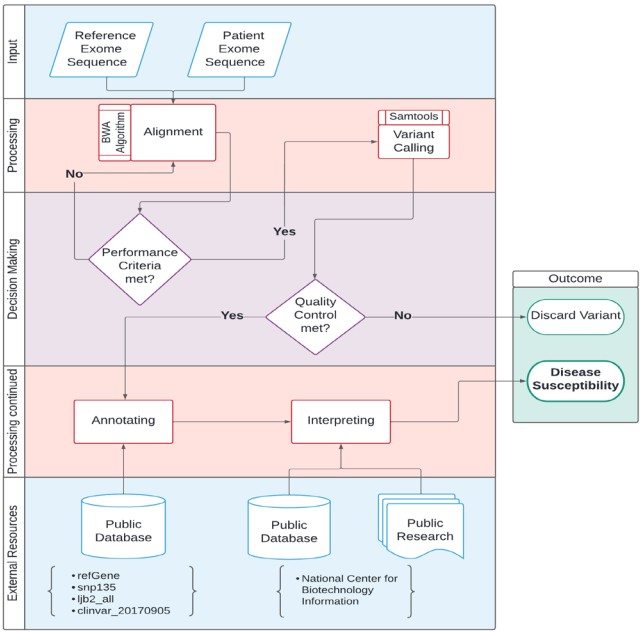

Figure 2: Research procedure. Flowchart depicting the logical steps within the stages of the exome sequencing process used in this project. The patient and reference exome sequences were obtained from the 1000 Genomes Project publicly available database.

on the assertion that exome sequencing methods can help identify genetic variants, including rare variants, that make people prone to certain diseases such as colorectal cancer.

By analyzing non-synonymous SNVs in the patient's exome sequence compared to the reference exome sequence, we hypothesize that it is possible to identify where the mutation occurs in the patient's exome sequence and detect the disease the patient might have a higher genetic risk for. Early detection and diagnosis can help improve survival rates.

RESULTS

In this project, we followed the exome sequencing process, which applies the Burrows Wheeler Alignment (BWA) Algorithm to identify SNVs and determine their role in human diseases (**Figure 2**). The BWA algorithm is widely used for mapping sequence reads to genomic databases because of its consistency and speed (26). The performance criteria established in this process maximized alignment between the reference exome sequence and the patient exome sequence by choosing the alignment read with the fewest differences between the two sequences, achieved with repeated runs of the BWA Algorithm. Subsequently, in this design, the results were measured against these performance criteria. After applying the BWA Algorithm, Samtools was used in variant calling and in determining the Phred score used for filtration of the variants.

During the exome sequencing procedure, we found 188,031 variants, of which 167,811 passed quality control (Phred score of 30). In addition, 7,256 synonymous SNVs and 6,107 non-synonymous SNVs (including missense and nonsense mutations) were identified. The distribution of synonymous SNVs was directly proportional to the number of

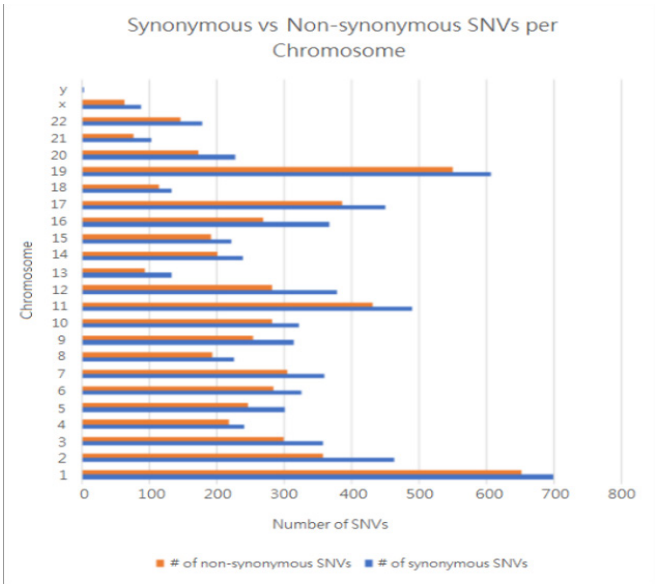

Figure 3: Synonymous vs non-synonymous single nucleotide variants (SNVs) per chromosome. Double bar graph depicting the number of non-synonymous SNVs and synonymous SNVs per this project's patient's chromosome.

non-synonymous SNVs in each chromosome (**Figure 3**). The number of synonymous SNVs was found to be greater than the number of non-synonymous SNVs for each chromosome.

Although our hypothesis was specific to colorectal cancer, to demonstrate that exome sequencing methods can be used to identify other genetic variants, this project was expanded to detect whether the patient had a higher genetic risk for diseases other than colorectal cancer. The program was redesigned to assert whether the performance criteria was met. A re-execution of the program ran a flagstat command to determine the accuracy of the optimal alignment read between the patient exome sequence and the reference exome sequence. The project found that of the 17,052,008 alignment reads that were processed, only 0.1% were rejected due to their inability to meet the quality control score of 30. This value of 0.1% met the performance criteria, showing a high confidence in the quality of the alignment process, and therefore a high confidence in detecting other diseases for which the patient may have a high genetic risk.

DISCUSSION

During annotation, one of the variants found was a nucleotide change from guanine to adenine, producing a missense non-synonymous SNV in Chromosome 5 at position 177093242 wherein the amino acid glycine was changed to arginine (rsID 351855). This non-synonymous SNV in the fibroblast growth factor receptor 4 (FGFR4) gene is known to promote accelerated cancer progression and tumor cell motility (27). One study suggests that the rsID 351855 variant (Gly388Arg) in FGFR4, combined with the gene's level of expression, affects colorectal cancer progression thereby supporting our hypothesis (28). The patient was also found to have other variants that show an increased genetic risk for colorectal cancer (**Table 1**) (29).

According to the National Library of Medicine, many Asians have been found to have the rsID 351855 mutation

Location of Variant	Type of Variant
Chromosome 1 Position 54780362 rsID 12144319 Gene: TTC22	3 Prime UTR Variant
Chromosome 2 Position 159108040 rsID 448513 Gene: TANC1	Intron Variant
Chromosome 3 Position 133982275 rsID 10049390 Gene: SLCO2A1	Intron Variant
Chromosome 4 Position 105207603 rsID 1391441 Gene: TET2	Intron Variant
Chromosome 9 Position 22103184 rsID 1537372 Gene: CDKN2B-AS1	Intron Variant
Chromosome 12 Position 57139907 rsID 4759277 Gene: LRP1	Intron Variant
Chromosome 13 Position 36887873 rsID 7333607 Gene: SMAD9	Intron Variant
Chromosome 15 Position 67110486 rsID 56324967 Gene: SMAD3	Intron Variant
Chromosome 17 Position 10803924 rsID 1078643 Gene: TMEM238L	Missense Variant
Chromosome 19 Position 58567729 rsID 73068325 Gene: MZF1	Intron Variant
Chromosome 20 Position 44037835 rsID6031311 Gene: TOX2	Intron Variant

Table 1: A partial list of variants found in the patient exome sequence that relate to an increased genetic risk for colorectal cancer

(**Figure 4**). Contrary to popular belief of colorectal cancer being most prevalent in Western countries, the incidence rates of colorectal cancer in Asia are high, and there is an increasing trend in the Asian population (30). Another study shows that in 2018, Asia accounted for nearly 52% of all colorectal cancer deaths worldwide which is due to an increase in its prevalence and incidence in Asia (31).

To better ensure that low-quality or low-confidence reads don't accidentally make it into downstream analysis, the quality control score threshold should be increased in future analyses. The procedure then identified SNVs and annotated

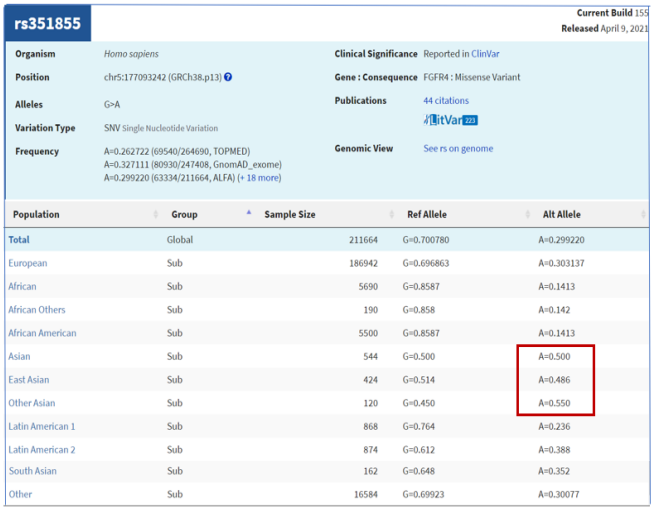

Figure 4. Frequency of the rsID 351855 mutation in ethnic groups across the world. The red box represents the frequency of this mutation in the Asian population sub-groups. (Retrieved from the National Center for Biotechnology Information - National Center for Biotechnology Information, 2021).

Location of Variant	Type of Variant	Associated Diseases
Chromosome 1, Position 100206504 rsID 1661800993 Gene: DBT	Non-synonymous SNV (deletion)	Type 2 intermediate maple syrup urine disease.
Chromosome 2, Position 227028004 rsID 2229813 Gene: COL4A4	Non-synonymous SNV (C > G)	Alport syndrome
Chromosome 5, Position 74685445 rsID 820878 Gene: HEXB	Non-synonymous SNV (T > C)	Hypomyelinating leukodystrophy
Chromosome 5, Position 177093242 rsID 351855 Gene: FGFR4	Non-synonymous SNV (G > A)	Cancer progression and tumor cell motility
Chromosome 7, Position 150999023 rsID 1799983 Gene: NOS3	Non-synonymous SNV (T > A)	Susceptibility to coronary artery spasm, susceptibility to late-onset Alzheimer's disease
Chromosome 9, Position 133436862 rsID 2301612 Gene: ADAMTS13	Non-synonymous SNV (C > A)	Upshaw-Schulman syndrome (type of thrombotic thrombocytopenic purpura)
Chromosome 11, Position 36593525 rsID 35691292 Gene: RAG2	Non-synonymous SNV (G > T)	Omenn syndrome (causes severe combined immunodeficiency)
Chromosome 16, Position 3254158 rsID 75977701 Gene: MEFV	Non-synonymous SNV (C > T)	Mediterranean fever
Chromosome 16, Position 27344882 rsID 1805010 Gene: IL4R	Non-synonymous SNV (A > G)	Atopy, resistance to human immunodeficiency virus type 1.
Chromosome 19, Position 12899706 rs8012 Gene: GCDH	Non-synonymous SNV (A > T)	Type 1 glutaric aciduria, glutaric acidemia
Chromosome 20, Position 4699605 rsID 1799990 Gene: PRNP	Non-synonymous SNV (A > G)	Early Alzheimer's disease, susceptibility to prion disease

Table 2: A partial list of the diseases for which the patient has a higher genetic risk other than colorectal cancer.

them with the biologically pertinent information found in the public databases (**Table 2**).

The annotating stage of the procedure during the retest found a missense non-synonymous SNV in Chromosome 16 at position 27344882: a nucleotide change from adenine to thymine changed the resulting amino acid from isoleucine to phenylalanine. This non-synonymous SNV in the interleukin 4 receptor (IL4R) gene (rsID 1805010) is known to cause atopy (32), with one study suggesting that this variant corresponds to an increased risk of asthma (33). Therefore, it can be interpreted that the patient has a higher genetic risk to asthma as well.

Based on the results of the non-synonymous SNV in the FGFR4 gene (rsID 351855) in the patient's exome sequence, which has been known to cause colorectal cancer, it can be concluded that the patient has a higher genetic risk for colorectal cancer. This finding shows that close monitoring of the colorectal area for any cancerous activity is recommended in order to detect colorectal cancer early and implement the best treatment plan possible thus increasing the chances of survival for this patient.

The performance criteria outlined in this procedure to determine the best possible alignment between the reference and patient exome sequences was met through the redesign and retest. This thereby increased the confidence in identifying whether the patient has a high genetic risk for colorectal cancer, and in identifying the exact mutation in the patient's exome sequence. It also led to the finding that the patient has a higher genetic risk for asthma due to the presence of the rsID 1805010 variant in the IL4R protein. In the future, this procedure can be applied with higher computational power to identify colorectal cancer in the Asian population given high incidence rates of colorectal cancer in the region. More broadly, this method can also be applied in detecting other diseases for which patients may have a high genetic risk.

While exome sequencing can be incredibly useful, the ability of clinics to process and store the data produced remains a significant challenge (34). This shows that exome sequencing must be improved to increase its use in a clinical setting. Another limitation of exome sequencing relates to the analytical validity of measuring the variants (35). It is important for a nucleotide to be read several times to be confident that the nucleotide was called correctly, thus implementing redundancy and preventing errors.

While exome sequencing is commonly the final diagnostic step in clinical genetics, it may miss diagnoses. In one study, overall, 36/54 (67%) of total diagnoses were based on clinical findings and coding variants found by exome sequencing while 18/54 (33%) of diagnoses were not solved exclusively by exome sequencing. Several methods were needed to detect and/or confirm the functional effects of the variants missed by exome sequencing, including genome sequencing (36).

Our study suggests that when specific gene testing panels do not provide a clear answer, genome sequencing should be considered before exome sequencing when available, because genome sequencing has increased coverage and diagnostic yield.

In this project, the patient's exome sequence was aligned with the reference exome sequence through sequence alignment to identify regions of similarity. A mutation in the patient's FGFR4 gene was discovered, which is known to promote rapid cancer growth and tumor cell motility. Future research into the clinical applications of exome sequencing should work to address the limitations mentioned above. With further research and repeated testing, exome sequencing is a promising method that can be applied to detect diseases more accurately and earlier on.

MATERIALS AND METHODS

The exome sequences were obtained from the publicly available database of the 1000 Genomes Project – an international collaboration aimed at comprehensively detailing the extent of human genetic variation by elucidating the entire genome sequences of several thousand individuals of various ethnicities from around the world. The patient exome sequence sample was obtained from B-Lymphocytes in the blood of a female of Vietnamese ethnicity. The age of the patient was not available, and the health of the patient was unknown.

The University of Chicago's supercomputer, Midway, was used to perform computational analysis. From this alignment, deviations from the reference exome sequence were determined using the BWA algorithm. To determine the best possible alignment between the patient's exome sequence and the reference exome sequence, the BWA algorithm entails repeatedly aligning the two sequences and computing a score for each alignment based on the differences between the sequences. The alignment with the least number of differences were then chosen as the alignment for the remainder of the project (37).

Exome Sequence Alignment

To facilitate genotyping and variant calling, the patient's aligned exome sequence was sorted according to the genomic coordinates of the reference exome sequence. The

patient exome sequence and reference exome sequence were aligned utilizing Samtools BWA algorithm, the most employed algorithm for exome sequence mapping (26), under the Samtools parameters "-q 5 -t 28" (38). Although there is a chance of missing true genomic variants while using the BWA algorithm, it is the most accurate method even though it takes longer when compared to other methods such as the Bowtie program (26).

Variant Calling

Next, we performed variant calling and identified both synonymous and non-synonymous SNVs in the patient's exome sequence compared to the reference exome sequence using the optimal alignment determined previously. Through this process, it is possible to enumerate the genotype of the patient's exome sequence at every position where it deviates from the reference exome sequence. The aligned sequences were then genotyped using the Samtools mpileup function, under the parameters "-t SP -uv -f" (38). This step computes the chromosome number, the mutation index of the chromosome, the nucleotide of the reference exome sequence, the nucleotide of the patient exome sequence (the mutation), and the Phred quality score reflecting the probability of the variant call.

Quality Control

Next, these variants went through quality control to filter the data to variants that might be worth studying later on in the project. Quality control is one of the most common metrics for assessing sequencing data quality. These data were filtered based on the Phred quality control score, which is modeled by the equation Q = -10*log10P, where Q represents the quality control score inputted, and P represents the probability of an incorrect base call. Variants with a quality control score of at least 30 were selected, at which point nearly all of the reads were perfect, having zero errors and ambiguities (39). This means that in selecting a quality score of 30, the base call accuracy of this project was 99.9%.

Annotating

After filtering through quality control, variants were annotated with biologically pertinent information:

1. The mutation type (e.g. synonymous, non-synonymous, etc.)
2. The RefGene ID of the affected gene
3. The dbSNP reference "rs" ID of the variant if it has been discovered previously

By annotating the variants using the tool Annovar, it is easier to decide which variants might be worth investigating, as it can be determined whether the variant has previously been categorized in a genetic disease. The genomic coordinates of the variants can be annotated by automatically comparing it to various public databases containing information regarding the prevalence of the variant. The variants were annotated with information using the refGene, snp135, ljb2_all, and clinvar_20170905 databases.

Finally, the annotated variants were explored for their potential roles in human disease. From these databases, information on the frequency of the mutation and its clinical significance was gathered.

ACKNOWLEDGEMENTS

This project would not have been possible without Dr. Lynne Muhammad, Dr. Andrew Mauer-Oats and Ms. Anna Gallardo from Whitney M. Young Magnet High School for their support, advice, and insightful comments.

Received: July 8, 2022
Accepted: January 2, 2023
Published: August 24, 2023

REFERENCES

1. American Cancer Society. "Colorectal Cancer Survival Rates: Colorectal Cancer Prognosis." Colorectal Cancer Survival Rates | Colorectal Cancer Prognosis, 1 Mar. 2023, www.cancer.org/cancer/types/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html.
2. Cancer.Net. "Colorectal Cancer - Statistics." *Cancer.Net*, 31 May 2022, www.cancer.net/cancer-types/colorectal-cancer/statistics.
3. Rawla, Prashanth et al. "Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors." *Przeglad gastroenterologiczny* vol. 14, no. 2, 6 Jan. 2019, pp. 89-103. doi:10.5114/pg.2018.81072.
4. Xi, Yue, and Pengfei Xu. "Global Colorectal Cancer Burden in 2020 and Projections to 2040." *Translational Oncology*, vol. 14, no. 10, Oct. 2021, pp. 101–174. doi:10.1016/j.tranon.2021.101174.
5. Colon Cancer Coalition. "Colon Cancer Facts." *Colon Cancer Coalition*, 28 Mar. 2022, coloncancercoalition.org/get-educated/what-you-need-to-know/colon-cancer-facts/#:~:text=In%202020%2C%20there%20will%20be,under%20the%20recommended%20screening%20age.
6. Ebell, Mark H., et al. "Cancer Screening Recommendations: An International Comparison of High Income Countries." *Public Health Reviews*, vol. 39, no. 1, 2 Mar. 2018. doi:10.1186/s40985-018-0080-0.
7. American Cancer Society. "Colorectal Cancer Guideline: How Often to Have Screening Tests." Colorectal Cancer Guideline | How Often to Have Screening Tests, 17 Nov. 2020, www.cancer.org/cancer/types/colon-rectal-cancer/detection-diagnosis-staging/acs-recommendations.html.
8. Cancer.Net. "Colorectal Cancer - Screening." *Cancer.Net*, 1 June 2022, www.cancer.net/cancer-types/colorectal-cancer/screening.
9. Healthline Editorial Team. "What You Should Know about Colonic Polyps." *Healthline*, Healthline Media, 4 Sept. 2019, www.healthline.com/health/colorectal-polyps.
10. American Cancer Society. "Colorectal Cancer Screening Tests: Sigmoidoscopy & Colonoscopy." *American Cancer Society*, 29 June 2020, www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/screening-tests-used.html.
11. Scripps. "Colonoscopy vs Sigmoidoscopy." *Scripps Health*, 22 Mar. 2022, www.scripps.org/news_items/4457-what-is-the-difference-between-a-colonoscopy-and-a-sigmoidoscopy.
12. Mayo Clinic Staff. "Virtual Colonoscopy." *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 23 July 2021, mayoclinic.org/tests-procedures/virtual-colonoscopy/about/pac-20385156.
13. National Cancer Institute. "Advances in Colorectal Cancer

Research." *National Cancer Institute*, 29 June 2021, www.cancer.gov/types/colorectal/research.
14. Illumina. *Genetic Analysis Education*, www.illumina.com/science/education/genetic-analysis.html.
15. National Human Genome Research Institute. "DNA Sequencing Fact Sheet." *Genome.gov*, 16 Aug. 2020, www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet.
16. Ranganathan Ganakammal, Satishkumar, and Emil Alexov. "An Ensemble Approach to Predict the Pathogenicity of Synonymous Variants." *Genes*, vol. 11, no. 9, 21 Sep. 2020, p. 1102. doi:10.3390/genes11091102.
17. Katsonis, Panagiotis, et al. "Single Nucleotide Variations: Biological Impact and Theoretical Interpretation." *Protein Science*, vol. 23, no. 12, 20 Oct. 2014, pp. 1650–1666. doi:10.1002/pro.2552.
18. U.S. National Library of Medicine. "What Are Proteins and What Do They Do?: Medlineplus Genetics." *MedlinePlus*, U.S. National Library of Medicine, 26 Mar. 2021, medlineplus.gov/genetics/understanding/howgeneswork/protein/.
19. National Human Genome Research Institute. "The Cost of Sequencing a Human Genome." *Genome.gov*, 1 Nov. 2021, www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.
20. Eisenstadt, Leah. "What Is Exome Sequencing?" *Broad Institute*, 24 June 2016, www.broadinstitute.org/blog/what-exome-sequencing.
21. Armaghany, Tannaz, et al. "Genetic Alterations in Colorectal Cancer." *Gastrointestinal Cancer Research*, International Society of Gastrointestinal Oncology, vol. 5, no. 1, Jan 2012, pp. 19-27. www.ncbi.nlm.nih.gov/pmc/articles/PMC3348713/
22. Zhang, Kejin. "Genetic Variations in Colorectal Cancer Risk and Clinical Outcome." *World Journal of Gastroenterology*, vol. 20, no. 15, 21 Apr. 2014, pp. 4167-4177. doi:10.3748/wjg.v20.i15.4167.
23. Miao, Beiping, et al. "Whole-Exome Sequencing Identifies a Novel Germline Variant in PTK7 Gene in Familial Colorectal Cancer." *International Journal of Molecular Sciences*, vol. 23, no. 3, 1 Jan. 2022, pp. 1295. doi:10.3390/ijms23031295.
24. Bjørklund, Sunniva Stordal, et al. "Widespread Alternative Exon Usage in Clinically Distinct Subtypes of Invasive Ductal Carcinoma." *Nature News*, Nature Publishing Group, 17 July 2017, www.nature.com/articles/s41598-017-05537-0?cid=tw&p.
25. Zhang, Jun, et al. "The Impact of Next-Generation Sequencing on Genomics." *Journal of Genetics and Genomics = Yi Chuan Xue Bao*, U.S. National Library of Medicine, 20 Mar. 2011, www.ncbi.nlm.nih.gov/pmc/articles/PMC3076108/.
26. Li, Heng, and Richard Durbin. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics* (Oxford, England), U.S. National Library of Medicine, 15 July 2009, www.ncbi.nlm.nih.gov/pmc/articles/PMC2705234/.
27. National Center for Biotechnology Information. "RS351855 RefSNP Report - DbSNP - NCBI." *National Center for Biotechnology Information*, U.S. National Library of Medicine, 9 Apr. 2021, www.ncbi.nlm.nih.gov/snp/rs351855?vertical_tab=true#frequency_tab.
28. Shiu, Bei-Hao, et al. "Impact of FGFR4 Gene Polymorphism on the Progression of Colorectal Cancer." *Diagnostics (Basel, Switzerland)*, U.S. National Library of Medicine, 28 May 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC8227855/.
29. Huyghe, Jeroen R, et al. "Discovery of Common and Rare Genetic Risk Variants for Colorectal Cancer." *Nature Genetics*, U.S. National Library of Medicine, Jan. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6358437.
30. Deng, Yanhong. "Rectal Cancer in Asian vs. Western Countries: Why the Variation in Incidence?" *Current Treatment Options in Oncology*, U.S. National Library of Medicine, 25 Sept. 2017, pubmed.ncbi.nlm.nih.gov/28948490/#:~:text=Colorectal%20cancer %20%28CRC%29%20is%20the%20third%20most%20common,Asian%20population.%20Furthermore%2C%20colorectal%20cancer%20accounts%20for%20%E2%80%A6.
31. Wong, Martin Cs, et al. "Prevalence and Risk Factors of Colorectal Cancer in Asia." *Intestinal Research*, Korean Association for the Study of Intestinal Diseases, July 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6667372/.
32. National Center for Biotechnology Information. "RS1805010 Refsnp Report - DbSNP - NCBI." *National Center for Biotechnology Information*, U.S. National Library of Medicine, 9 Apr. 2021, www.ncbi.nlm.nih.gov/snp/rs1805010#frequency_tab.
33. Burgos, Paula I, et al. "Association of IL4R Single-Nucleotide Polymorphisms with Rheumatoid Nodules in African Americans with Rheumatoid Arthritis." *Arthritis Research & Therapy*, BioMed Central, 5 May 2010, www.ncbi.nlm.nih.gov/pmc/articles/PMC2911851/.
34. Bertier, Gabrielle, et al. "Unsolved Challenges of Clinical Whole-Exome Sequencing: A Systematic Literature Review of End-Users' Views." *BMC Medical Genomics*, BioMed Central, 11 Aug. 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC4982236/.
35. Merogenomics. "Advantages and Limitations of Genome Sequencing." *Advantages and Limitations of Genome Sequencing | Merogenomics Inc.*, merogenomics.ca/en/advantages-and-limitations-of-genome-sequencing/.
36. Burdick, Kendall J, et al. "Limitations of Exome Sequencing in Detecting Rare and Undiagnosed Diseases." *American Journal of Medical Genetics. Part A*, U.S. National Library of Medicine, 19 Mar. 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC8057342/.
37. bwa.1. "Manual Reference Pages - BWA (1)." *Bwa*.1, 8 Mar. 2013, bio-bwa.sourceforge.net/bwa.shtml.
38. Li, Heng, et al. *Samtools(1) Manual Page*, 22 Oct. 2021, www.htslib.org/doc/samtools.html.
39. Illumina. "Quality Scores for next-Generation Sequencing - Illumina." *Quality Scores for Next-Generation Sequencing*, 12 Mar. 2012, www.illumina.com/Documents/products/technotes/technote_Q-Scores.pdf.