Post Research Write-up

By: Michael Rogers

Note: Parts of this research have been redacted in the interest of non-disclosure.

Abstract: In the recent past, many people have been developing ways to detect and predict birth defects and physical disorders that happen at the chromosomal level in embryos. Many fertility clinics have found ways to predict many of the diseases that can cause miscarriages or crippling disorders when a child is born. In the fetus' DNA, there are 23 pairs of chromosomes that determine so many different parts of a human. In this research project, myself and a team of very talented individuals worked together to create a process that can detect the copy number variation in a section of the human genome and determine if a fetus will be healthy or will have some of the terrible diseases that come along with a gain or loss in chromosomes. In order to accomplish this task, we used a python library called CNVkit, which is linked to in the references of this report. We found that when you use this software to compare a section of a patient genome to the genome of a healthy chromosomal make-up, it makes things vey clear about how healthy the fetus will be. Not only is this information extremely important to have, but the information must also be visualized in a way that is easy for the doctors to read.

Motivating problem: In the interest of finding viable embryos, finding the gains and losses in chromosomes is incredibly important to knowing if an embryo is going to have a disease or will be healthy. Copies in chromosomes may seem like a somewhat insignificant problem when we think about all of the things that can go wrong in the process of making a child. However, copies or losses in a person's genome can affect the way a person lives and can also cause death before the fetus is even able to make it into the world. An example of some of these gains or losses in chromosomes is called Trisomy 21 commonly referred to as Down Syndrome. Trisomy 21 is one of the only gains or losses that a person can survive. It is when there is a duplicate of the 21st chromosome in a person's genome. Even though an embryo can survive it still inflicts both a physical and mental toll on the person living with it. The motivation of this project is to be able to detect these abnormalities and choose how to proceed with the pregnancy. This comes along with both moral and religious arguments for and against this process. Even though all of these factors play into this problem, how can we possibly know what were doing if we can't read the data? This problem comes down to how well we can visualize this data and understand what it means.

The Findings: The goal of our team was to be able to provide the software and method in order to compile a copy number variation report that shows the gains in losses in specific chromosomes. This report will not dive into the exact process that allows these charts to be made, but it will go over the science and results of running the raw data through this method. Each dataset is originally given in a .BAM format. There is extensive cleaning that must be done on the data before any kind of report can be compiled or visualized. This is because the .BAM file is strings of DNA proteins. The sample data from the embryo is then tested against a human genome that is known to have one of each chromosome. The sample used to obtain the visualization in figure 1 is the only sample in this report that will be discussed. From the third party that we are doing this research for, we know that this patient has a gain on chromosome 7 and a loss on chromosome 8.

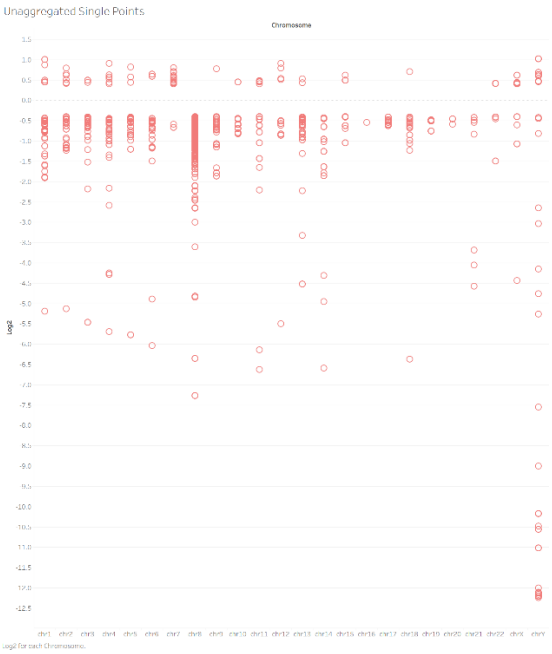

Figure 1 – Raw data of gains and losses visualized.

As you can see, this char is nearly impossible to read and have a definitive answer as to if there is a gain or a loss on a chromosome because of the sheer number of points. After putting the data through out process we were left with a .txt file that needed to be aggregated because of how messy the data was on its face. An interesting thing to notice about

figure 2 is the supposed loss in chromosome Y. This is because the subject used is a female.

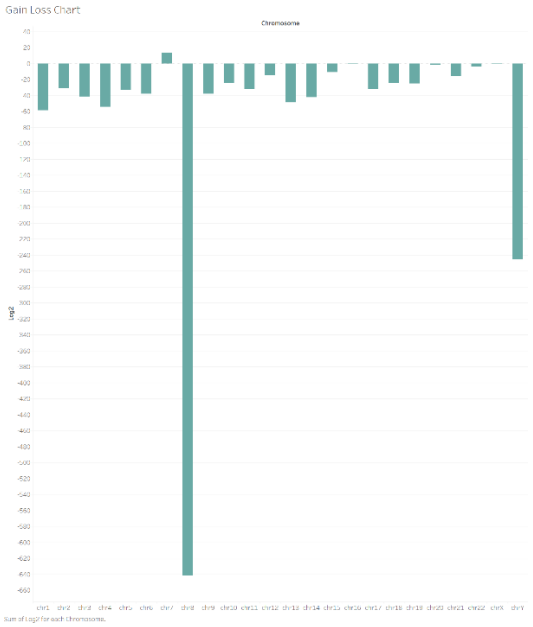

Figure 2 – Aggregated raw data to get a clearer view of gains and losses.

As seen in figure two, we can see much more clearly what chromosome has a gain and which has a loss. It is very obvious that after we aggregated the points associated with each chromosome. Although this graph is much easier to read, it helps to have the raw data at hand and also be able to see the aggregation without actually aggregating it. This can be seen in figure 3.

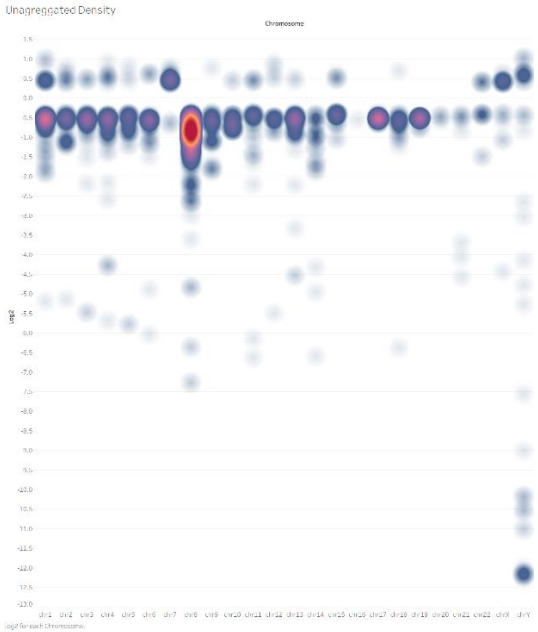

Figure 3 – Unaggregated density chart.

As you can see from figure 3, we have an idea of the representation of the raw data that has been aggregated—but not aggregated—to get the effect of figure 1 and 2 combined together. When these three visualizations are looked at together, the chart from figure 1 begins to look far less confusing and almost adds to insight into what's happening with the data.

Moral Argument: Although this process seems to be a very important and helpful piece of technology, there is a legitimate and interesting argument to be made that by use of this process we could end up discarding an embryo that could grow to be a person like Steven Hawking. This is a very interesting argument, because we just don't know what all diseases are caused by these gains and losses in chromosomes. There are also those who would make the argument that by using this process, we are "playing god" in picking viable fetuses. This is also an interesting argument because at what point does picking and choosing become bioengineering? There are so many different areas of study that this problem quickly becomes extremely relevant. It is applicable in medicine, philosophy, computer science, genetics and so many others.

Conclusion: In conclusion, this project was so interesting to be apart of. There are so many implications and applications of this technology. Visualizations are the first step of these applications because if we can't visualize these gains and losses in chromosomes, then how can we possibly progress? If there is no way to read the data, then it would make all of our efforts to get the copy number variation useless. We can't use just a single visualization to tell the entire story. If we only use a representation like the one showed in figure 2, then we don't have the whole story. But when the data is presented in multiple different ways, then it makes it easier to understand and progresses technology quicker than

Citations:

[1] "Genome-wide copy number from high-throughput sequencing," *CNVkit*. [Online]. Available: https://cnvkit.readthedocs.io/en/stable/. [Accessed: 14-Dec-2019].

[2] "Genome-wide copy number from high-throughput sequencing," *CNVkit*. [Online]. Available: https://cnvkit.readthedocs.io/en/stable/. [Accessed: 14-Dec-2019].

[3] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavaré, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis, "Relative impact of nucleotide and copy number variation on gene expression phenotypes," *Science (New York, N.Y.)*, 09-Feb-

2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2665772/. [Accessed: 14-Dec-2019].

[4] *Nature News*. [Online]. Available: https://www.nature.com/scitable/topicpage/copy-number-variation-445/. [Accessed: 14-Dec-2019].

[5] J. L. Freeman, G. H. Perry, L. Feuk, R. Redon, S. A. McCarroll, D. M. Altshuler, H. Aburatani, K. W. Jones, C. Tyler-Smith, M. E. Hurles, N. P. Carter, S. W. Scherer, and C. Lee, "Copy number variation: New insights in genome diversity," *Genome Research*, 01-Jan-1970. [Online]. Available: https://genome.cshlp.org/content/16/8/949.short. [Accessed: 14-Dec-2019].

[6] "Copy number calling pipeline¶," *Copy number calling pipeline - CNVkit 0.9.5 documentation*. [Online]. Available: https://cnvkit.readthedocs.io/en/stable/pipeline.html. [Accessed: 14-Dec-2019].

[7] "What is copy number variation? ." [Online]. Available: https://www.gene-quantification.de/cnv-faq.pdf.

[8] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. MacArthur, J. R. MacDonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles, "Origins and functional impact of copy number variation in the human genome," *Nature News*, 07-Oct-2009. [Online]. Available: https://www.nature.com/articles/nature08516. [Accessed: 14-Dec-2019].