Dan Geraghty, a researcher at Fred Hutchinson Cancer Research Center and CEO of Scisco Genetics, has spent much of his career focused on the genetics of immune response. Recently he talked to Mendelspod host Theral Timpson as part of a series of podcasts on the rise of long-read sequencing.

Geraghty explained that while there have been decades’ worth of studies associating the genetics of the major histocompatibility complex (MHC), and the highly polymorphic HLA class 1 and 2 genes, we still haven’t found the key mutations for a variety of different autoimmune diseases such as type 1 diabetes, rheumatoid arthritis, multiple sclerosis, and others.

Enormous amounts of linkage disequilibrium in these regions are one factor, as is getting information in phase, so larger stretches of sequence are needed. Recently Geraghty has begun using Single Molecule, Real-Time (SMRT®) Technology with hopes of drilling down to the causal genetics.

**The Challenge With Short Reads**

Geraghty explained that sequencing fosmids with short-read technology is cumbersome when it comes to stitching together the reads. Data analysis and finishing “became a roadblock that the Illumina short-read technology wouldn’t let us get beyond,” he said, noting that the finishing process takes 30 minutes to an hour per fosmid, prohibitive for any modest-scale effort. Geraghty marveled that he has received 40 kb reads from PacBio® sequencing – meaning a whole fosmid can be sequenced in one piece.

**PacBio Is Ready to Handle the Challenge**

Geraghty said that with recent technology improvements, PacBio data is “really high quality” and “as good or better than Illumina and Sanger,” noting that his group has compared all three technologies with the same sequences. “It opens up a whole new possibility,” he said, because previously “you simply weren’t getting all of the data. People were using statistics to impute missing data and so on, and it simply hasn’t worked.”

Geraghty thinks so, noting that a resource such as the 1000 Genomes Project would be upgraded significantly with PacBio data for complex regions such as MHC and KIR. He said that if you look at these regions in the 1000 Genomes data you will find “a mass of confusion” because those regions are highly repetitive and contain a large amount of copy number and allelic variation, making it difficult or impossible to assemble the data correctly with short reads.

“Any large human genome sequencing projects using only short-read technology are not going to acquire usable data for these complex regions, it’s as simple as that,” he said. For complex regions, “you’ll need long-read data,” he said, “The long-read data will give you really what everybody has been after all along without realizing it. It will give you the phase and all the detail on the polymorphism in these highly polymorphic regions.”

**The Future Is Bright**

Geraghty expressed his excitement about the future using long-read sequencing this way: “We’re hot on the trail. We basically see the entire picture; we are not looking under a lamp post for the keys. It’s daylight and we can see the whole neighborhood. So we’re going to find the keys.”

*The full interview can be found at www.Mendelspod.com*