**The G Word Transcript**

**Prof Matthew Hurles: The Sherlock Holmes moment**

**Chris:** Hi, I'm Chris Wigley, CEO of Genomics England. I've spent my career at the intersection of technology, ethics and human stories. Now, I lead the amazing team here at Genomics England. We're trying to bring the benefits of genomic medicine to everyone, and that involves accelerating genomic research, and also working with the NHS to bring genomics into the heart of healthcare. Genomics is a word that can trigger really strong responses, hope, fear, anger, and there's a lot of information out there. But it's not all accessible to non-experts, and there are some myths out there. So, we want to talk more about this word, The G Word - genomics. That's what this podcast is about. Welcome to The G Word. It's my huge pleasure to welcome Professor Matt Hurles onto the podcast today. He is the Head of Human Genetics at the Wellcome Sanger Institute and Senior Group Leader of the Hurles Group, which is obviously named after him, as well as being an honorary professor of Human Genetics and Genomics at the University of Cambridge, and the co-founder of Congenica, which is a big diagnostic decision support platform. Matt, welcome to the pod.

**Matt:** Thank you very much, Chris. It's a great honour, and thanks for inviting me.

**Chris:** If we can start with, you're now leaving this big group at Wellcome, you're a Fellow of the Royal Society, a storied leader in the genomics field. Tell us a bit about how five-year-old Matt became Matt Hurles that we see today. How did you get into the field? And how did you get excited about it, and so on?

**Matt:** I guess I didn't really know what I wanted to do. So, for most of my school career, I kind of hedged my bets and took the options that least require me to narrow down what I can potentially do in the future. So, I ended up doing biochemistry at university because it was a very broad subject. And it allowed me to go in lots of different directions potentially. And I just really discovered through actually an extended project during my undergraduates that I was really interested in the genetics side, and actually genetics and evolution initially. And that really stimulated me to go and do a PhD, which wasn't something I was necessarily expecting to do. And that was very much in the how we use genetics to understand human prehistory. So, it was much more about population genetics than about medical genetics.

**Chris:** Wow. So, this is sort of all the way back to how do we get from kind of great apes to humans? Or more like cave people? What sort of period or scope?

**Matt:** Yeah, so actually, population genetics, evolutionary genetics, allows you to go everything from what happened in the last few 1000 years to way back to the beginning of life. The bit that I was most focused on was how did humans get to where they got to in the world, and how we can use genetics as a tool, as a record of the past to trace these prehistoric migrations. During my PhD, I focussed most on the prehistoric migrations of the Polynesians through the Pacific and how you can trace using genetic lineages, which islands they got to, when, and in what order.

**Chris:** Wow, that is incredibly cool. I spent a number of years of my childhood in New Zealand in the far north, it's a very cosmopolitan mix of Polynesian island. Amazing, fascinating subject. And so, from the PhD, you largely continued in an academic path, how has that played out? If I'm a student today, wanting to get into the field, how should I think about it?

**Matt:** Well, certainly, finding something that you're passionate about, and then finding a supervisor who you really get on with, who you think's going to invest in you, it's often much more about the individual than it is about the topic. I was quite lucky to find the good combination. A guy called Mark Jobling in Leicester, who was pretty young at the time, I think I was a second PhD student, but gave me a fantastic training, really invested and looked after me during that, really formative period of your scientific career. And so, I got a real passion for tracing prehistoric migration as something I wanted to do after my PhD and wanted to continue with it. So, I actually did a postdoc in an archaeological research institute, led by a very forward-looking archaeologist called Colin Renfrew, who really thought that archaeologists and geneticists should be in the same building, talking the same language as much as possible. And so, I did a postdoc there.

**Chris:** I have a mental image of you as Indiana Jones in a dusty desert somewhere fighting the Nazis. Is that what it was like?

**Matt:** No, definitely not. It was me in a lab coat talking to the people who were actually doing that over coffee and thinking what a much more exotic life they lead. During that postdoc, I really stumbled across a much more biomedically focused insight. And I really stumbled across it, because we were using the human Y chromosome to trace lineages through the Pacific. And it was just as the human genome sequence was starting to come out and a portion of the human Y chromosome came out. And I realised that this bit of the Y chromosome that I was looking at, that was telling me about a particular type of migration in a particular part of the world, was also critically important for male fertility. And if you deleted that part of the Y chromosome, a male can no longer produce sperm. And that insight about how a rare variant can have a dramatic effect on a person's life and how one can use population genetics training, to make insights in medical genetics was for me a really pivotal moment. I was just then very fortunate, because I was clearly very much in the wrong place there to take that forward. But I was very fortunate that Sanger Institute was at that point in time pivoting from being a genome sequencing factory to becoming a research institute, driven by its faculty research interests and somehow, I managed to sneak under the radar and join as a junior faculty.

**Chris:** Fantastic. You built on that biomedical insight into this whole programme around the Deciphering of Developmental Disorders programme, the DDD programme, which is an ancestor in some ways of the of the 100,000 Genomes Project, let's dive into the meat of this. You and your team are looking at genetic information, clinical information, other data sources, to try and really unpick these incredibly complex questions around how we as humans develop, how some of the quirks and some of those inputs can lead to different outcomes for us as kids, as humans? How's that work evolved over the last, I guess about 20 years, since the heavy lifting of the Human Genome Project? What do you and the team do day to day? I realised that's a huge scope of a question, but maybe take us by the hand of how this work actually gets done, and how that's changed over the last little while.

**Matt:** The key thing was that the genome was both an end in itself, but also a beginning for a whole technological revolution, because it was just one genome. And really, if we want to know something about differences between us, whether it's disease or not disease based, we really need to understand how the differences, the genetic differences, on that one genome, cause those different changes, why we're different from one another. And so, it's all about the genetic variation. And what we really needed after the genome was then cost-effective ways to identify the genetic variants that people carry. And the genome serves as a template and allows you to design those assays. So that's the key point. The first one off the block, was using microarray methods to detect very large deletions and duplications in individuals.

**Chris:** There's this basic point about the human genome with one genome. But if we're looking for, where is an individual difference, the one human who the human genome was based on, something that was in their genome might be in every genome, it might be a unique variant to them. So, we need that broader perspective about almost "what does normal look like?" Before you can start to assess whether some individual piece of the genome is normal or unusual or quirky in some way.

**Matt:** In reality, what tends to happen is that a new technology comes along, and a group of people very focused on patients with rare diseases, immediately wants to apply that technology and discovers a huge amount of new variation they couldn't previously see. In parallel with that, another group of people really wants to understand population variation around the world, and then characterising this general population variation. It's really the combination of those two types of data - "What does normal look like?", for want of a better phrase, and "what unusual genetic variants do patients with rare diseases carry?" It's the combination and the comparison of those two, that's really powerful. What was interesting to me was that, on one hand, one can take a dataset and do population genetics and understand mutation processes and selection, and these things that population geneticists care about. On the other hand, exactly the same technology and a different set of individuals, would say here is a clear cause of disease, this is going to be meaningful for the family for all the reasons that diagnoses are meaningful. One of the big challenges was when the first technology came along, which was the micro arrays, it created a new challenge for clinical geneticists, which was this bit of our genome is deleted. I don't really know much about this bit of the genome, because I'm now looking across the whole genome, and I've never really done that before at this scale, how can I find out if this bit of the genome is important and relevant to the disease, or is it just something that's just benign, and it's not relevant to disease. What that stimulated was the importance of taking genetic variation in data, aligning it with what we know about the genome, and what we see in healthy individuals. That was the genesis, that was actually Helen Firth and Nigel Carter at Sanger coming up with the idea of DECIPHER and create an international web portal, where clinicians could post on behalf of their patients 'I have this genetic variant, I think maybe the cause of disease, but I'm just not sure what tools can we provide that help the clinician interpret that and can we identify any patients anywhere else in the world that might have a similar genetic variant and a similar condition?' That's proven to be extremely powerful, and that kind of insight initially from looking at microarrays and large chunks of DNA that are gained and lost, has transitioned in exactly the same way into the next generation sequencing technology that came along about five years later and started to say, 'here's this patient who's got this rare, seemingly damaging variant in this gene that I don't really know much about, are there any other patients who have similar damaging variants in that gene?' That matching process internationally, has proven to be incredibly powerful with defining new conditions that couldn't be done just by looking at that patient in isolation.

**Chris:** Absolutely. Within that context, how did the cohort for Deciphering Developmental Disease come about? How did you pick people to go into that cohort and what kind of analyses have you done on them?

**Matt:** It started actually, as predominantly an array-based project to look at large CNVs. As the project was just starting off, that's when the next gen sequencing revolution came about, so we rapidly pivoted to doing exome sequencing, so sequencing all of the genes in the genome. That proved to be incredibly powerful. The insight that certainly Helen and Nigel had at the inception of the DDD study was that comparing the DNA of a child with a developmental disorder to their parents, from the outset, was going to be incredibly valuable and that collecting DNA from the parents at the same time was going to help hugely in the interpretation of the genetic variation you see in the child. That intuition, which came partly from the CNV work, from being able to identify a copy number variant that was present in a child, but was absent from either parent was hugely influential in interpreting whether that genetic variant was likely to be the cause of the child's disorder or not, and transitioning that into the next generation sequencing and what we now know is the trio based approach and we think of a standard in developmental disorders. That really came out of some of that early work in looking at those large deletions and duplications. From the outset of the project, we really wanted to engage the family, the child, their parents, collect DNA from them, and systematically collect phenotype data. Because one of the challenges that we had with the DECIPHER portal was that although one could submit genetic information and phenotypic information, getting systematic phenotypic information was really challenging. This is well in advance of the year that we now live in with respect to electronic health records.

**Chris:** But still tough, right?

**Matt:** Which is still tough. We were really fortunate with the clinicians driving the project, Helen Firth, David Fitzpatrick, working very closely with the hundreds of clinical geneticists in the UK and working out what was the phenotypic information that was really important was going to be really meaningful to capture, that could be captured on the ground by clinicians, that would sit alongside and enrich the genetic data. And that again, was their fundamental insight that really empowered the whole project and allowed it to evolve with the genomic technology.

**Chris:** Try and bring this to life for us a bit with, I don't know if there are any individual family stories that you can share, but what kind of conditions were picked up through the programme? Did people get referred to it because they presented in the NHS with a condition and people didn't know what to do? What kind of impact could a family see through taking part in the programme?

**Matt:** It's very much a partnership between the central team at the Sanger Institute and all of the clinical genetics’ centres in the UK and in the Republic of Ireland. Essentially, those clinical geneticists were caring for those families, thought they had a genetic condition, but couldn't find the genetic cause of that condition. But we're pretty confident it had an underlying genetic cause. We weren't going to be too restrictive at the outset about the phenotypic criteria, they had to be quite extreme, and therefore more likely to have a genetic cause. But we really trusted the clinical geneticists' intuition about 'is this likely to be genetics?' and that was a really good starting point for the project.

**Chris:** What kind of symptoms would that be? I don't know if there's a typical case, but is that things like seizures, or there’s this broad term of 'potential failure to thrive' but how would that present for a kid and their family?

**Matt:** About 85-90% of the cohort have neurodevelopmental disorders, so intellectual disability, or developmental delay if they're younger. The average age of recruitment to the project is about six and that's typically when a child has really started to struggle in the first year or two of school, even if parents have been able to detect it and be concerned earlier. That's often when they get through the clinical system to a clinical geneticist and clinical tests are done. That's typically the median age that we found was around six. You're right, a quarter of those children will have seizures, about 10% will have other organ malformations, heart malformations being the most common, but in amongst that there's actually going to be many 1000s of different conditions represented in the cohort. Actually, we were deliberately taking a 'all comers' approach with the idea with that maybe a condition might not look the same in two different children, maybe it might manifest it slightly differently in different children. Unless we take that broad approach, we won't capture that and that certainly turned out to be true.

**Chris:** I'm going to drop two quotes in from former The G Word podcast here. The first quote is what I love from Eric Topol, which is "every disease is a rare disease because it presents in you". I think this is very much the point you're making here about the same disease can manifest differently in different kids, depending on some combination of different glitches or quirks in their genetics, differences in environment and exposure, and so on. The second quote is from Euan Ashley in Stanford, who talks about the process of gathering clues and doing "detective work" like Sherlock Holmes to try and understand what is going on with this kid, or adult for that matter. Help us understand how you and the team gear up to do that Sherlock Holmes work, you've got clues in different places. How do you get to the "aha" moment?

**Matt:** One of the processes we've tried to do is maybe a little bit different from what's been done in some other centres where maybe they've had a smaller number of patients, they've invested a lot in in characterising those patients and then they undertake the Sherlock Holmes type work that you describe. We thought one of the biggest advantages of having a National Health Service is by casting that net widely about bringing together a really large cohort of families, and then developing new statistical methods that enable us to rather than have that in particular Sherlock Holmes insight into one particular family, we can actually say, using the power of statistics, 'here is a gene that we really ought to be paying much more close attention to. These are the families that look like they've got a damaging genetic variant in that gene,' and take this really strong genotype driven approach to identifying new conditions. Also challenge ourselves from the outset of the project in the same way that you guys at Genomics England have challenged yourself and say, 'let's make this work at scale diagnostically'. Let's take on the challenge of informatically really searching for the needle in a haystack in an individual's DNA. Let's do it in such a way that it can benefit 1000s of patients. That was something that we really stepped up to from the outset and certainly created sleepless nights and stresses and strains along the way. But ultimately, we think is very much in keeping with our ethos of the National Health Service about not just solving the one or two cases, but actually providing a service that actually works for a population.

**Chris:** That point about equity of access, and breadth of impact, I think is hugely inspiring. In that context, I'm guessing that in some clinical areas, you're seeing real insight at scale from these approaches. You say, 'we've cracked this area', in other areas, it's probably still more complex or unsolved. What kind of impact are we seeing in terms of the kinds of diagnostic rates that we see in different areas? What do you think is the next chapter for kids, adults, families who haven't yet had a primary diagnosis, but are still living their lives with those symptoms?

**Matt:** This is something we've been looking at really carefully, just recently, and Caroline Wright's been very much leading on this part of the team that I that I co-lead, it's very much a team work. What Caroline has been asking has been, what are the factors that influence whether we can diagnose somebody through DDD. In or out of that we, on average, we can diagnose about 35 to 40% of families that come through DDD, but clearly, we're not diagnosing up to two thirds of families. If we look at those factors, there's a number of really intriguing things in there. We do a better job at diagnosing girls than we do boys, we do a better job of diagnosing families when there's only one affected child than when there's two affected children, we do a better job of diagnosing families where we can get DNA from both parents from where we can't get DNA from both parents. We also do a worse job in diagnosing families with who have African ancestry. We can only tease these patterns out when you look across many 1000s of families, but it does really point towards the areas where we really need to push. How do we make diagnostics work better for single parent families? How do we make it better across different ethnicities? What are the challenges that we can tackle there? I think that really helps us know where to push on. From a scientific perspective, one of the things we can also do is say, 'let's take out all of the diagnoses in this large cohort and let's look at the remaining patients that we haven't diagnosed and let's compare them to a control population who don't have disorders, and let's say where does the damaging variation look like it's still remaining.' We can see that there's still a lot of damaging rare genetic variation in protein coding genes in those undiagnosed cases. In many cases, we have identified the variant that is the root cause of the child's condition, but we haven't yet gathered enough evidence that that particular gene is associated with disease. What that speaks to us is, what we really need to do, is increase the size of the populations beyond the 13,000 families that we work with in DDD, we need to combine the data with Genomics England, with collaborators outside the UK and only then can we start to gather enough evidence to say, 'yes, this gene really is associated.' There's that class of missing diagnoses, which is where we've discovered the variant, but we just don't have enough evidence that that gene's associated with disease.

**Chris:** In some cases, you might need to characterise a new disease almost. If we understand the genetic causes, we understand the set of symptoms, but it may or may not match to an existing disease definition almost. It's that almost inchoate space.

**Matt:** Yes, this is where the discovery process goes hand in hand with the diagnostic processes, you have to make those discoveries to make those new diagnoses. Another thing that the data is pointing towards is that there's, amongst those undiagnosed children, the role of damaging genetic variation that is inherited from a parent who is seemingly much less affected or not affected at all, to the degree that the child is really quite important. That's pointing us towards the idea that genetic variants in different contexts might have different severities of effect. We don't understand that nearly so much., it's a much more complex way of looking at genetics that is challenging both from a scientific perspective to understand but also from a clinical perspective of how to use that information.

**Chris:** Analytically, that's presumably extraordinarily complex to unpick because there may be some other variation in some other parts of the genome that may be countermanding the deleterious effect in the first part of the genome you're looking at. I'm guessing that that's where, as you mentioned, having more datasets to work across, but it's also a question of using more and more sophisticated analytic techniques over time.

**Matt:** Yeah, and more sophisticated analytic techniques and collecting more data on the family. For example, one of the things that we discovered in doing this diagnostic work, and especially shout out to Patrick Campbell, and Hillary Martin, who really drove this statistical analysis, was that children who are premature are less likely to get a genetic diagnosis. That suggests that there's an environmental contribution to their disorder, it could well be, for example, that the reason that a child has a damaging inherited variant but that has a much more severe phenotype than their parent, was because there was an in utero environmental trigger in combination with a genetic variant. That's an area that we really want to explore more. To get into that space, one really has to do a good job of characterising that in utero environment and using the clinical data that might be available on how that pregnancy progressed.

**Chris:** There's so much in what you've just said that I really want to unpick, let me try and pull in a couple of threads. One is the work that you mentioned around trying to understand the drivers of whether a given child gets a diagnosis or doesn't get a diagnosis. How are you and the team thinking about correlation and causation? For example, you mentioned a kid that's part of a trio where we have both parents as well, much more likely to get a diagnosis and that intuitively makes sense, we've got a fuller genetic picture to work with. You also mentioned that folks from African ancestral backgrounds were less likely to get a diagnosis, I'm making intuitive loops here but I think I'm right in saying that, from having read the paper a little while ago, that part of that is because people from African ancestral backgrounds are less likely to be part of a trio. Is that actually the causal factor? Or is it something in the genetics of people from an African ancestral background? How do you go about unpicking some of those quite nuanced questions about correlation versus causation?

**Matt:** Yes, that was a key part of actually what Patrick and Hillary did was to put all these factors together into one statistical model so that you can tease them out. What they could show is that, for example, if there's an African ancestry child who is recruited as a trio with DNA from both parents, the diagnostic success rate is the same as other ancestries. The challenge is really around when we can't get DNA from both parents. There's one fundamental population genetic reason for that is that individuals with African ancestry have more genetic variation, there is more genetic diversity in Africa. As a consequence, the haystack is slightly larger, but the needle is the same size. There's one fundamental population genetic challenge there. But there's an additional challenge that we can do something about, which I know you're passionate about this as well, is that the reference datasets of normal population variation in individuals of African ancestry are just smaller, and therefore less powerful than the ones of European ancestry. What we really need to do is enrich those African ancestry population datasets, and that's something we really can hit on and I know you guys plans to do so. I think that's a real global challenge.

**Chris:** You took the words out of my mouth without wanting to advertise our own work - the Diverse Data programme. Hugely important. I'd love your thoughts, you mentioned earlier about back in the very early noughties, quote unquote, 'the human genome' from the Human Genome Project, which is one genome. We're now in a position where we have hundreds of 1000s, millions of genomes that have been sequenced using next generation sequencing and so on. There are a number of reference genomes out there. You just mentioned the point about the data set that a quote unquote, 'African ancestry reference genome is based on being smaller, alongside the higher level of genetic diversity within there's different communities and groups in Africa, because of the nature of kind of human migration' . Where do we see this going? You go from one reference genome that you can compare in the new genome too, to saying, let's get to five reference genomes for major ancestral groups, East Asian, African and European whatever. Then you're like actually come on, there's no such thing as an African genome, there's show so populations, other populations, other communities. I'm making it up, let's say there's 20 different reference genomes in Africa, do we ultimately end up, and this is a bit of a nerdy question, but bear with me, with some kind of bespoke reference set for a given person's genome, so instead of having a reference genome, you say, this person is from this community, blah, blah, blah, but the relevant reference set for their genome is actually dynamic rather than static? Does that make sense as a question?

**Matt:** I think so, but as a population geneticist, one of the things about humans, as unusual about humans compared to other large body mammals, is that we actually have very little genetic diversity. We have a very recent common origin, predominantly in Africa, and only a subset of that diversity ever made it out of Africa. That means that one reference genome actually works pretty well across populations. I don't see the reference genome itself, the genome to which we aligned reads to, as being a major contributor to diagnosing currently undiagnosed individuals of African ancestry, which I think is fortunate, but is a property of our prehistory. It could have been that we had a very different prehistory, that we'd had a larger, more global population for a longer period of time and actually, maybe we really would have needed different reference genomes. I also think there are ways in which one can, technical ways, in which one can compile reference genomes from around the world into one graph reference and potentially use that. That has been a technology that has been hugely attractive and has been 18 months away from prime time for about 10 years. We're fortunate in a way that we do have this low genetic diversity means one reference genome, which is itself, the reference genome, is itself a compendium of individuals. It's not one individual and it's a compendium of individuals with different ancestries.

**Chris:** Back to the point about needing larger data sets on different populations. From your perspective, is that principally around effectively understanding the commonness or rareness of variation from the reference unit in a given population? Then it's about how do you define that population that you're assessing rareness or commonness within?

**Matt:** It's really about enriching those existing global genetic variation data resources for individuals that are special and we should be over sampling Africa, because of the greater genetic diversity, not under sampling it as we have done currently. But we're fortunate that we can sequence those individuals and map them on to a reference and get most of the value out of that using our standard reference. But it's really about developing those projects and partnerships that enable those products to be done in a trustworthy way such that individuals of African ancestry around the world can benefit.

**Chris:** Let me pull on a slightly different thread now from the work you talked about on the DDD cohort. You mentioned involving the families in the dialogue about this, trying to understand richer phenotypic data, other types of datasets. What have you learned from that work and other work about how researchers can best interact with families, with whole communities who have a given recognition, in a way which is useful and beneficial to the families and useful and beneficial to the researchers? Obviously, the research then hopefully ends up benefiting the families as well. How can we get those kinds of engagement programmes right?

**Matt:** That's a really good question. Certainly, being a research project and DDD, all the families are completely anonymous to me and my research team. We really work through the clinical community to have that interaction which is a bit different from the engagement that you've done with many of the participants much more directly. That is very inspiring what you guys have done. We did a really good job at the time of working out what data to collect on families and what clinicians could reasonably provide in a period of time that enabled us to do a large-scale study. But it wasn't the ideal. The ideal would be something that involves both record linkage, especially to healthcare records to enable us to get a much fuller picture. For example, in DDD, we have very light phenotypic information on the parents of children, so if one does identify a variant that's inherited from a parent and looks like it might be relevant in the child, it's very limited what we can do in terms of following that up in terms of was there any unusual developmental history of that parent that might support the idea that that variant is actually playing a role in the child. Whereas in the 100,000 Genomes Project, because you have worked with the participants to do health record linkage equally well in the parents and the children, that's now gives us an opportunity that in my group we're now working on at the moment, which is to actually interrogate that question at the next level of granularity using that linked healthcare data. The other thing that we did was working with SWAN UK and unique patient support and advocacy groups at the time to represent the interests of the participants was really helpful. You got a real sense there of patients and families being the real experts of their condition, but not really at the time having the mechanism by which they could contribute to enriching the phenotypic data, double checking the phenotypic data, and making sure it really represents their child and their family in the most accurate way. That's a real opportunity and I know something that you're exploring, and I'm very keen to help with that and thinking through how can we allow families to contribute and effectively be a participant in the research process, and increase the probability that they'll get a useful answer for their family out of it.

**Chris:** It's a fascinating area, a couple of thoughts strike me. I'm spit balling here, but one is that we know that electronic health records are not great, in terms of accuracy, depth, etc., which is no criticism of the health system, or of those platforms. NHS Digital ran a pilot in Leeds a little while ago where people with the NHS app popped up and said, 'would you like to check that your phone number and address and email address are correct in NHS records?' Most of the people who got the pop up, clicked on it, and most of them then made a change to that information. That's just your name, email address, phone number and home address, let alone the details of a specific element about how frequently do you have seizures or if you have a short left leg, how short is it compared to your other leg, is there a foot on it, does the foot have toes? These much more granular pieces of phenotypic information. It does feel like there's a huge opportunity to do that in dialogue. It's interesting to see what some organisations in the private sector have done like 23andMe around engaging with people who use that app. As far as I understand, those communities are super willing to really invest the time and thought to make that as rich a dataset as possible. How do you think in, I was about to say the real world, in the world of academia, the NHS, public policy in the UK, we could best do that practically?

**Matt:** There's a number of bits where families are already providing information to the healthcare system, it's just not cascading through into these research datasets. One example that I think is really interesting is the Red Book, which we have on our three kids, the Red Book, where you track the growth of your child and various other things.

**Chris:** It's almost as if it's got a sacred status in families.

**Matt:** Exactly, but I understand that's moving to a digitised form very shortly and certainly in DDD, we found that collecting that measurement information was actually hugely helpful. We could identify certain new disorders, because there was a very unusual growth pattern, for example, in the children. We collected that in DDD, it took quite a lot of effort to collect that, but the advent of a digital Red Book means that that should be a additional piece of information that the families are already providing to the health service, which would cascade through with the right connections and linkages into the research. There's that kind of opportunity. There's also the fact that NHS Digital only has a very high-level view of the health data available in hospitals and hospitals in EHRs often have a much more granular view and finding mechanisms by which we can have that more granular data come through into the research department, I think is really important. The third strand is of course that engagement, direct engagement, and having the platforms that families want to engage with that give them information back, but also enable them to contribute more information. That combination of deeper linkages new ways of bringing information in from the healthcare system and families contributing will create amazing datasets that allow us to answer questions we have in our minds now but also questions we don't have in our minds. I can give you one really good example of a question that I don't think anyone had in their mind that we just recently managed to publish and work our way through using the 100,000 Genomes Project data, which is that we identified the small number of families where children had a very unusually large number of new mutations that their parents didn't have. The DNA is copied incredibly well from generation to generation, but one in every 50 million letters of DNA is erroneously copied in everyone. These children had about rather than one 50 million had more like one in 10 million, so it was maybe a five times higher mutation rate, which is incredibly unusual. Just to be clear, there's a tiny number of families about 11 families out of the 13,000 trios in the Genomics England dataset that we looked at that have this property. It's a very rare thing. But what we were able to do was, for a number of those families, about five of those families, we could look back into the parental health data and see that the father had been treated with a chemotherapy, typically due to a cancer prior to the child being conceived. We could also link up the chemotherapy that was used, which, with a particular mutational signature, that we could see in the child. I don't think anyone at the outset who thought, let's do health care linkage, let's make sure we do it for the parents as well as the children, was thinking about the effects of chemotherapy potentially on new mutations in a child. That's the key value of these general datasets as they allow us to ask completely new unthought of questions.

**Chris:** Wow. That's an absolutely fascinating story. We're already dealing with an extraordinary amount of complexity here. 3.2 billion base pairs in DNA, all of the complexities we've just been talking about, about clinical data where it resides, deep phenotypical data. Just blow our minds a little bit with transcriptomics, proteomics, metabolomics. How are you in the team thinking about those other molecular datasets, imaging datasets? Is this just going to explode all of our minds? How can we use it productively to get to more diagnosis, more patient impact? How far along on the arc from fundamental research to clinical application do you think we are?

**Matt:** Looking purely through a diagnostic lens, there's two lenses one can look at this, one is through a mechanistic lens, how can we use these additional layers of data to understand the mechanism by which the genetic variant gives rise to the condition potentially think, therefore, about where there might be opportunities to intervene therapeutically. But through a purely diagnostic lens, the big challenge that we have is that we don't know what most variants do, we don't know if they're damaging or not, even in genes, which we are absolutely confident, are associated with disease. For example, there are about 1500 genes that are associated with developmental disorders that are robustly associated that are used diagnostically. Only about 10% of the variance in those genes have ever been seen in a human before, so 90% have never been seen in human before. Most of the diagnoses we make in the DDD study are in that 90% of variance we've never seen before. The big challenge we have is how well can we distinguish between variants that do cause a disease that fall in the gene and variants that don't. We know that most variants don't cause disease already. We can see plenty of people out there and the general population who have variants in those genes that don't cause disease. It's not just I see a variant in a gene there and our genes associated with disease, therefore, it must be the cause. This is known as the variant of uncertain significance problem. It's a big blocker on the genomic medicine in general. There's two ways in which one can think about using these datasets to solve that. One is to, for a given disorder, for a given disorder caused by damaging mutations in a gene, to define some multi omics signature that is characteristic of that disorder. That could be an unusual pattern of RNA, it could be an unusual pattern of DNA methylation. There's about over 50 neurodevelopmental disorder genes now, where there's a disease specific pattern of DNA methylation in blood, so not in brain but in blood that we can test. That can be diagnostically very useful to characterise a sample from a patient who has a variant of uncertain significance and be able to say 'yes, it is or isn't damaging'. That's that approach, which is the reactive approach, I've seen a variant in a patient, therefore I do another assay. The thing that's really exciting me at the moment is the prospect of doing those kinds of assays in cells, where one can make every single possible mutation in a gene, and read out whether it's damaging or not, and so have proactively determined which are the variants that are likely to be diagnostically relevant, and which are likely to be benign. This is a new technology that's just coming along and this field is very much in its infancy but it feels like a Human Genome Project in the making, it feels like the very early stages of the Human Genome Project where we know that fundamentally this is the information we need to be able to interpret every variant of the genome, which is where we want to get to. It's going to get us there quicker than sequencing patients even over decades over all countries in the world. But we have a technology which currently doesn't enable us to scale, we can do one gene or two genes, we need to do 4000 genes.

**Chris:** Unpick that for us a tiny bit, how are you in the team exploring that? How do you see that being deployed?

**Matt:** In terms of a deployment, it fits well within the existing frameworks by which clinicians interpret genetic variation. There is a one can use functional data as it's called to up weight or down weight a genetic variant. What we're talking here is about systematically generating that functional data. In many respects a lot of the practice framework is in place, what's not in place is the technical ability to generate those data and disseminate them to all the places that needs to go to every clinician's decision support interface around the world. What we're doing at the moment is building a community of researchers and companies who are infused by this vision. It's called the Atlas of Variant Effects, the thing that we want to build and generate that community and expand that community and enable that community to collectively generate these maps of variant effects and every single disease gene over the next 5 to 10 years.

**Chris:** It's super exciting, I'm conscious we're close up on time, but maybe a couple of questions to look to the future. One is, at Genomics England we're in the process of putting together this programme around newborn sequencing alongside the heel prick test, for anyone listening just Google Genomics England Newborns and the whole thing will pop up. If that all goes according to plan, which we hope it will, we will have a relationship with 100,000 families who have chosen to take part in the study, who understand what's involved and whose baby will have their genome sequenced at birth to flag potential issues that affect kids in early life, create an amazing data set for research into those conditions for patient benefit. There will also be the first cohort of kids to be born with a genome on file. What for you are the implications of that? How might we see that helping to advance these insights over almost the next generation?

**Matt:** I and many people who think about this, think this is the end state. In 50 years’ time, people will have their genome at birth that will be used clinically, potentially non clinically, people find value in that. One of the challenges thinking about how do we get from where we are now to that position and the thing that's so exciting when I think about the newborns programme, is it's one of the few routes by which you can see that happening. It is not just a narrow focus on the 'how does this augment the newborn screening that we're already doing?'. But there's also the forward-looking thinking about 'how might this information be used in the future? How can we make it available to a broader research community who can bring that future into being?' There's only going to be a small number of these initiatives that are really catalysing. What will then happen is other countries who want to undertake genomic medicine programmes nationwide, cover multiple domains of the use of genomic information, will think, 'we need to do this as well, let’s see how this would work in our context.' I'm hugely enthused because it's been easy to say 'everyone will have their genome at birth in some decades time'. It'd be hard to plot a route by how we get there. A route that has an ongoing relationship with participants and has a trustworthy way of engaging the research community to think about how can we get the value out of the genetic information that's useful for the healthcare service, but also is useful for the participants themselves, what information do they want? How do they want to use their information? One of the things I'm acutely aware of is that we're often diagnosing children with neurodevelopmental disorders at the age of six or later and many of those parents will be thinking, 'had I had that information earlier, what might we have done that might have improved the outcomes for my child?' It's very hard to see how we would ever answer that question for them without doing a programme such as this.

**Chris:** Hugely inspiring, we need to leave it here, we may have to get you back on to the pod for a follow up conversation at some point, because there's just too many interesting topics to explore with you. In the meantime, a final quickfire question is part of the rationale behind us having this podcast and having these conversations is to try and have more of a national conversation around genomics as it's coming more into the mainstream. Are there any themes or any individual people who you think we don't hear enough about or from? Who else should we get onto the pod?

**Matt:** It's maybe a bit challenging but I think there's some really interesting work being done on how genetic variation influences child outcomes in a more educational sense than in a clinical sense. It's a contentious area, but it is something that is well supported by the data, and we have to engage with as a society. The risk is not engaging with where the data and research is taking us. We can see, for example, there are genetic variants that are clearly diagnoses for patients in 100,000 Genomes Project that we also see in people in the general population, in population cohorts, such as UK Biobank. We can clearly see in populations such as UK Biobank those genetic variants are clearly having an influence on the outcomes of those individuals. Currently, we're looking through two very different lenses, we're looking at this clinical lens and this population lens. In this era, 50 years hence, where everyone's got their genome, that will be one lens. We need to have a set of conversations over a period of time that starts to bring those two lenses together.

**Chris:** Hugely inspiring. Learning more and more about how we work, how we can each be living our best lives. This is the mission ahead of us. Challenge accepted. Matt, thank you so much for joining us. Incredibly inspiring to talk to you as always and thanks again.

**Matt:** It's been a great pleasure, Chris. I'm delighted that you're A doing this podcast and B have had so much positive feedback and a big audience because it's really important.

**Chris:** That's all for this episode. Thanks for listening to this discussion about The G Word and for joining us on this journey to highlight and debate the implications of genomics as it comes to the mainstream of healthcare and society. Remember to subscribe to The G Word on Apple podcasts, Spotify, or wherever you listen. If you have views on these topics, if you have a suggestion for someone we should interview, then do write to us at podcast@genomicsengland.co.uk. Do remember if you've enjoyed listening that giving us a five-star review really helps other people find out about the series, I'd appreciate it very much. See you on the next episode of The G Word.