Tokyo Workshop on Genomic Medi...eutics, and Health 20240411-04

Thu, Apr 11, 2024 11:49AM • 26:05

**SUMMARY KEYWORDS**

data, sequencing, released, uk biobank, samples, genotyping, participants, pipeline, project, cohort, datasets, analysis, genetic, uk, perform, single cell sequencing, download, quality control, researchers, imputation

00:20

So I'm gonna give a quick overview of the many and various bits of genomic data that we have. I know that's a key interest to a lot of people in the room who might be wanting to use that exclusively or in combination with the other health records and participants information. So just for a fast look DynoMax has formed a huge part of the timeline of UK Biobank from the end of recruitment. It was only a short number of years before the genotyping data was released. The project to start the genotyping started almost immediately after recruitment phases that had ended. I think it was actually kind of the speed at which genomic technologies have increased when recruitment began. It was thought it would be impossible even to genotype participants at that scale. That started as a pilot of just 10% of the cohort. And very quickly at Stanford to the full title bullet. Only a year after the genotyping data was released, when I realized that it was possible to do whole genome sequencing on the full cohort and that study began a year later. So in just 10 years, we've gone from not being a pipe dream to that being achievable and released to the research community. So there's four main data sets for the or dataset types for the genetic data across scales. So the original data was the genotyping that was followed up with imputation data. These are both available on the whole cohort. Then there's a whole exome sequencing which was released shortly after the end of the genotyping project, which is on about 95% of the cohort. And there's a couple of different processing pipelines available for that. And then whole genome sequencing which is on 490,000 participants, which was released at the end of last year. We've made the decision recently that genomic scale data is no longer available for download this is available through our cloud platform only. So the whole exome or whole genome sequencing data is now only available on the cloud and recent releases of imputation so want to know exactly when and what my top meds which were released. Within the last year. I've also only been available free so there's no further download access for people who have older products, projects, which was started before about 2021 They may have a downloaded copy of the original release all that same sequencing. So up to the first 200,000 was once downloaded. And project to have continued I got to keep those on premises copies, the genotyping data and the original imputation is still available for download for projects in tier two and above. So the genotyping data was the first genomic data released this was done using a custom built alpha matrix genotyping array, which looked at a 500,000 variants. So the first 50,000 participants formed a pilot and they were done in the UK believe axiom. And then the remainder were done on the UK Biobank axiom array which was specifically designed for that from that trucks in the UK Biobank population. There's about 95% overlap between the two arrays. And the data was made available for all the participants in 2017. There's some additional data which was released alongside the genotyping which is really important to understand it and also can be useful in its own right. So there's information about relatedness as I mentioned earlier, that cohort is highly interrelated as people who have 10s of relatives with

05:38

Tell us which facts the data was not quite. That's important, for instance, for identifying which participants are in that first pilot, which are on a slightly separate survey, which you'll need to potentially consider definitely within your analysis. We have plank and be done versions of the thing should be made available. Then there was an original phasing and imputation, which was released within a year of the file genotyping release. This was done by the Wellcome Trust Center for Human Genetics and Oxford, and they perform these analyses using smoke.

06:18

This was done using the haplotype reference consortium and

06:21

UK 10k as reference sets are very UK based populations. There was also implantation of the HLA alleles. It's important to note that this original imputation genotyping array data is on gr ch 37. The remainder of genetic datasets on doing docx 38. So these are you know, you can't directly compare between the two without performing leftover or similar. There are also two additional imputed datasets which were released in 2002. So these are both in form invitations based on the original genotyping array data. One was performed using genomics England as a Reference panel so they have 100,000 Genomes Project and are based in the UK and the other was off the top note two aspects, which is from 54 parts and

07:23

approximately diversity notes are designed to represent lots of different

07:28

ethnic groups. These are both designed to perform well in more diverse cohorts. So depending on the population, which you're interested in, if you're using amputation data, you may want to consider the merits of the difficulty paddles, the coverage and the reasons that you're looking at learning literature about how well they perform, particularly if you're interested in specific subsets of the pillbox, as opposed to that. We also then have full access to sequencing. So this is sequencing of all the coding regions of the genome. This was done using human genomes commercial funding, and was sequenced on a number of nodes see that for the first 50,000 sequences released in 2019. And then there was a final release of the whole cohort of 70,000 in mid 2022. And there's been some really interesting results coming out of this. It's shown that you can do some deep analyses that's accounting you visit products. And it's a really popular data set because it's got a huge amount of power. It's a huge amount of information in it, and it's a lot smaller and slightly less unwieldy than the audience. So then, just looking back at this timeline, giving us a talk about the whole genome desktop, just to show the kind of time that project has taken which is important to consider or looking at the data which has come out about our first sample the sequence and this project on the 30th of August 2018. And then the last four years went up nearly in 2022, with the data released publicly, largely consulted privately in 2023 at the beginning and released publicly at the end of 2023. So this is a sort of five year project or some important kind of quality data to consider as you go through this and as you use the data. So one of the most important is that the sequencing was shot between two providers. So the data was in two tranches. We've had a Vanguard phase, which is a pilot. That was 50,000 participants who are sequenced by wellcome Sanger Institute in Cambridge in the UK, and processed by some pretzels with your partner. The remaining 450,000 samples were evenly split between Depot dynetics, Iceland. The Welcome Centers together. So sample sequence by decode will process with an in house pipeline at decode genetics, sample sequence by Soma either as part of the pilot or the main phase through a process by seven bridges, and they used pipelines which covered all the same QC lab tracks all the same concepts, but they weren't completely identical. There was small changes as well between the Vanguard's and the main phase pipeline for several samples as they updated certain elements, improved performance, as also important to consider this although we have samples which was shipped as Vanguard or processed using that pipeline, that the fast nine phase sample was sequence just as the 50,000 and finished processing. The last Vanguard sample was sequenced about three months before the end of the project. How samples were shipped that designation in terms of flights isn't the only way to identify the differences between them. So it's really important to consider a separate pipeline to us withdrawal report and the data. So the samples are safe and sometimes quite late data resequencing of case to ensure that they have fast processing so to help with that there's QC and additional data. So there's procedural and quality control data available for or sequence. So that includes the factors they were sent in. So how they were shipped, which provider they went to which pipelines they were used to analyze those samples. And they're really important to consider in your dataset, particularly if you're interested in say rare variants or small numbers of participants who may have particularly how they are compared to the rest of the population. There's also information from the quality control that was before. So things like the contamination levels, which are identified and the samples that were sequence, and genotype order, then the structural barrier datasets as well which have been produced for the dragon underground type of data. CMA STR data, which is available. Currently, we have face data available for

12:59

the interim graph type release.

13:01

So that was the first 200,000 participants who were released at the end of 2020. And that was done using the vehicle pipeline and also said the flights are done by two different research groups. And it's been really interesting to see the kind of comparisons and performance and outcome those threats to data in combination we're making available, so stay tuned because that's the final draft of the title. So the final full genome sequencing data is released into law firms. So there's the web man slash do it pay slack because it's a continuous pipeline that lots of bits batch together, and that's what we've used for the original interim releases. Some basil pipelines. It's used widely within research already. And then there was also an additional bathroom, which was made available using the dragon pipeline. So that sort of single button starts. And that's only available for the final beta release. So whole genome sequencing data because it can't be taken off the graph. We consider that really important to ensure that was a person that vailable which could enable more analysis across datasets by national analysis. So this would allow us to have datasets producing similarly as possible to other airports. So the dragon ipipeline, because it's all of us. And that means that there's something ready to go which is as similar as possible to this

14:50

because we're aware that realignment or recall in sequencing

14:54

data is just not feasible to the majority researchers, both time experience, cost, that kind of work. But within the current community, the GTK and graph type pipelines are more well known. They're making more problems on a day to using our cousin's house. And so making both fashions available means hopefully cover research needs for all groups. So there's a few other bits of genetic data which are sometimes more overlooked slightly, so I thought I would put a little bit of an emphasis on some of those additional genetic datasets which we've made available. So leukocyte to mele was an analysis produced by the University of Nottingham, which was done on residual genotyping samples after genotyping and then use the sample sites when there was no residual remaining. And then they measured that in the full cohort that was released in 2001. So a normalized standardized length divided by they've done some really nice work. You can see that they do correlate very well with agency and it starts and also quite a big gap last night, seven year difference in aged between men and women, which is interesting to say. There's also some PRs scores, which were generated from the genotyping data by DynoMax PLC. So they've won PRs for 28 common conditions and 25 but it's not the price. They've also made available a comparison tokens that's like no doubt at the bottom called UK B. Pratt, which allows a comparison of that PRs scores to any others which you might generate using your own algorithms or algorithms to have a standardized way of comparing

17:08

performance of PRs algorithms.

17:11

Because these are done largely on diagnosis. As they said, You have a lot of data on diagnoses and conditions are not updated regularly or continuously seeing the doctor and dying hospital or having surgeries. So the intention is that this data will be updated regularly. And so diversity to the first update will be released in June of this year to account for new diagnoses of tenses. So we have some our upcoming genetic data so there is a clonal hematopoiesis data that this has been is a piece of research that has been a lot of different groups and slightly different datasets. We're making available one that was on the 450,000 participants for XM sequencing by a group who also combined in combine the data with all of us combined analysis of multiple races, which is quite interesting to see and that should be available towards the end of this year. There's also a research group in the UK who has been using UK Biobank samples for integrated human herpes viruses. So this is a direct analysis of UK

18:41

Biobank samples for mostly integrated

18:44

virus sex using fast attack plan assay, but they've also expanded to doing some long read sequencing to get more information about the exact lineages for those participants that I can find that will make the available data very shortly during this year, and the final data set which is included normally the sequencing of those reasons, is likely to be a favorable towards the end of this year or early next year. And then there's a couple of other genetics projects which are in the pipeline and one of those is single cell RNA sequencing. So there was a pileup of similar stuff taking place in the next couple that was done by a call from Antonius group at wellcome Sanger Institute. This has been a really hot project pipeline, they require me to fresh samples, which meant every day as approximately 60 participants attended out or anything for next. Located about 300 miles apart across the UK. Those samples are then couriered overnight to arrive in Cambridge within 24 hours the processor. So it's been a project which has been a real learning curve in terms of how you perform this kind of single cell sequencing at scale and effectively today 5000 participants who attend tonight's over about an eight month period, and that did include some of those neuro Tenzin COVID imaging study. So it will be really interesting when not least because it will include we won't have the prior time point obviously because this has only been done recently. Our samples will that we will have this case control participants and to see differences in the single cell sequencing advice, so that data is currently in preparation. It's estimated fewer, it's not clear, made available

20:56

to the research community about time. Really, really interesting to see what people make that and suddenly we're continuing to do single cell sequencing at larger scale

21:12

Yeah, so as public pilots are considering as well, but yeah. Okay, I think that is a whistlestop tour of

21:24

all of the genetics data. So if there's any questions I'm more than happy to take them

21:46

apologies if it's too specific, but since you mentioned the single cell RNA seek agenda since it's hokey BMC probably how many cells capturing per per donor

21:56

roughly was their target? Now you're stressing me operates participants in the same sequencing we have

22:07

in combination as well as doing single cell sequencing.

22:10

We also did a whole choir constraint what happens on those participants, which will be released in combination doesn't have itself

22:18

because that's great. If you have a

22:22

whole blood count data then it was more to think of which cell types be represented, if any are

22:26

in the minority, but if there's a blood count to be scaled up, identified cell types will be released. In addition to the kind of quality control outcomes thanks. Question.

22:51

Does the follow up question. So why not just about our receipt first, law causes large samples. It's something worth considering and quite a thing at the moment. It's something Daisy mentioned in her presentation this morning about how you access the data is other than the Clinical Biochemistry which was informed by UK Biobank. These studies we do are driven by the researchers who come to us with a proposal

23:27

with content and so that means that that determines

23:33

the order in which things are carried out. UK Biobank as an organization, we support those approved applications, but we don't generally go for specific research funding for specific grant

23:48

outside projects, generally of our own.

23:52

There's a few small exceptions. Along the way. Mostly it's

23:57

made by Greece research is important to us.

24:01

Okay, so another question that having said that, you mentioned that this data is recorded.

24:08

Yes, that's not our reporting platform. It's created by genomic scale See, available, as a suggested for researchers who want to look at the performance process. Yeah, I think it's very useful to have you guys consider about launching some computational challenge. We're finally going to dream challenge where you release a set of training sets and

24:34

researchers develop their method. Now you you've either submitted reductionism asset, it's something we've considered receptors have created application to the UK Biobank to do developments of computational methods as opposed to sort of classic

24:54

medical research. And that is an accepted method for an application as well. If you want to access the data to develop methods for others to use.

25:12

Big question, you know with data access, understanding, especially the raw data is only accessible through the platform. Learn parts of the data like the PCs

25:24

that can be downloaded or No, no, I said that genotyping aside, which, because

25:32

of legacy on SAS, we continue to allow for the latest data releases. Only auxilary

25:41

data is available for download. So for instance, you could download the quality control metrics for WDS. If you want, for instance, that is offline. You can't download any of the files which contain sequencing because we'll come back to that at the end of the day. Combine data Costco or something. Yeah, it's I think it's in common when you're releasing data on a scale of questions.

26:18

Right on time, so we have an hour break now.