Iain Kennedy's FullGenomes (and other) results

FGC5856

May 5th, 2019

FGC5856, FGC5864, A5599 branches

My analysis results from the sequencing of my Y-chromosome were delivered by Dr. Greg Magoon of FullGenomes Inc. (FGC) on November 22nd 2013, following the return of the sequenced data from BGI in Hong Kong. The raw data itself, an 8.1Gb BAM file, was downloaded and shared with my research team on December 13th.

The core objective of the test was to locate new variants in my DNA below my current terminal ISOGG SNP which in my case is R1b-M222. The analysis does however cover my genetic branch above R1b-M222 and some of this latter information may also be of use to researchers, particularly those studying the SNPs just above M222 such as Z2961 and DF13.

The basic variant types included in the analysis are SNPs, INDELs and STRs. For now I will pass over the latter although they are quite numerous as they don't particularly interest me. The SNP and INDEL reports are further sub-divided into 'public' and 'private' sections. These terms are commonly used but continue to cause confusion in the genetic genealogy arena since everyone seems to have a slightly different interpretation of what a 'private' variant is. In the case of the FGC analysis they specifically differentiate between variants also seen in public domain genomes and those that aren't; note that the analysis does not make reference to other FGC customers and their data. There has been some criticism of this but I think FGC has taken the correct route.

The variants are further sub-divided by quality. Each segment has been read 50 times during sequencing but the quality of the data still varies and FGC have created four levels of data quality. These are not comparable across all the reports and in some instances a result for a SNP the customer is already known to be derived for has a poor quality indicator.

In theory then my report should show 16 different categories of variant (2 variant types x 2 privacy types x 4 quality levels). To save time for now I will put to one side the 3rd and 4th quality levels; the top two levels are supposed to represent 99% and 95% certain call values and should be good enough and there seems to be a big drop down to the next quality level, making these variants of questionable use.

So how many have I got? Based on the variantCompare spreadsheet,

SNPs: shared=154, private=22
INDELS: shared=30, private=1

high quality (95-99% confidence) variants.

There have been some scientific findings that a Y SNP mutates about once every 90 years on average. Using this figure would make R1b-M222 90x22=1980 years old, or 2070 years old if we include the INDEL too. I will not dwell on this calculated figure since we already know from other sequencing candidates that each of us has had a different number of SNPs and dating calculations are not the purpose of the exercise from my point of view; the objective is a rigorously defined unambiguous Y SNP tree.

The list of shared SNPs is not of course all the SNPs I share with public domain data as this analysis has been largely limited to a part of the R haplogroup.

Another small point is that two of my 22 private SNPs may not actually be private but their exact position in the tree is still in some doubt.

In order to get these variants into service they need to be evaluated and hopefully trialled. Another geneticist from outside FGC has now vetted all my 23 private variants and concluded that in his opinion only 15 are likely to be individually testable. The commonest reason for

rejection for the other 8 was them being in one of the highly repetitive regions of the Y chromosome. These are now being prepared so that they can be used for testing against other people. In addition and more importantly for now, my two lowest level shared SNPs which were only shared with other M222 genomes have also been prepared and are being tried out as of now. The ideal outcome is that these two latter SNPs define the remaining undefined sub-branch of M222, currently reported as accounting for some 17% of testers. It is possible that I am not in said branch (if indeed it is a single branch) but in a small early branch with much more limited membership and only time will tell.

What is the overall testing strategy now? For the time being my private SNPs go hand in hand with the Chromo2 chip which should put 83% of M222 into a branch of their own; testers who fail to be placed in a branch and who are reported M222* can then test against my SNPs to explore this part of the tree further.

Update on YFull.com (Dec 29, 2013).

In December of 2013 I shared my raw sequencing data with another DNA analysis firm YFull.com. This has resulted in the unveiling of a further 9 SNPs not found by FullGenomes.com (although there were also a similar number they missed so the accuracy rates of the two analysts are probably similar). This means that my total to date is actually 32 SNPs not 23 and hence the tentative dating calculations above are out. Once again Thomas Krahn has analysed the findings and reduced the usable number of new SNPs from 9 down to 5 for the same reasons as with the original list - either the variant was in a highly repetitive region or it also appeared in other chromosomes. Since I have plenty of SNPs for evaluation I am trusting Dr. Krahn although its possible in the future I might re-examine some of the discarded ones if I find someone else who shared many of my private SNPs.The new variants are labelled YFS* by the analysis company and range from YFS041777 to YFS041880.

Meanwhile another person has tested positive for FGC4077 and FGC4078 and is now exploring the lower level private SNPs; and two other people reported as S7073* by Chromo2 are trying FGC4077/8.

Anniversary update (Nov 14, 2014)

It is now a full year since I received my sequencing results from FullGenomes Inc. Another FGC customer, Mr. Wilson, has turned up in the FGC4077/8 branch although he didn't match any lower level SNPs so represents a different sub-branch. This was no less a figure than the person who discovered the M222 haplotype and made the connection between it and the M222 SNP, and the first non-Kennedy DNA match I corresponded with back in 2004. He has been followed by a number of other sequencing testers. These did a reduced coverage test which didn't actually include FGC4077/8 but fortunately a third marker FGC4087 which so far appears equivalent was included in their test. Follow up custom SNP testing by further people has demonstrated that all three SNPs are at the same position in the tree and it is probably sufficient to only test one of the trio (we call this 'phylogenetic equivalence').

My own Kennedy testing has concentrated on a core group of four including myself who are all resident in Scotland and all trace to Perthshire or Angus. All four of us are now confirmed to at least match FGC5856, one of the private FGC SNPs found in my test. One has been fully tested against all 20 of my markers and mismatches me on 6. Similar testing will commence shortly on the other two. It is hoped that a mini-tree structure can be worked out, however experience from others suggests that the SNPs may occur in small clumps rather than one at a time so how far we can further break down that group of six markers remains to be seen. There are several non-Kennedys who are potentially in my sub-branch, their surnames include Bruce, Robertson and Gow. The total proven membership of the FGC4077 group who have agreed to go public is 31, and they are illustrated on the M222 chart I maintain.

Additional branching has been found under FGC4077, with branches marked by FGC12948, A725 and A360. Of these, early signs are that A725 is the most important. The geographical spread of the group is virtually the same as that of M222 as a whole, from the west coast of Ireland to the east coast of Scotland. Extrapolating back from surnames whose age is << 1000y to determine an origin for these markers is a dubious scientific exercise. One of the two people directly above us ie those who are S7073+ and FGC4077-, S658-, S568- has a Manx surname and the other has a trail to Donegal but a surname that might be argued could be Irish or

Scottish. In some ways FGC4077 is like a parallel version of M222 with S660 taking over as what we once thought of as M222. This means progress is going to be hard to achieve as most research is geared towards the main part of the group ie S660 and its sub-branches (now up to seven in number). It remains to be proven whether FGC5856 reaches outside the Scottish highlands. Branches are continually dying out so what we have left may not be an accurate guide to how the tree evolved.

All the markers referred to above are currently available on the YSEQ 'DF49&M222' panel which is priced at just USD88 and tests all known sub-branches of M222. The full set of FGC5856 branch markers are also available as a package at YSEQ although not as cheaply since there is less chance of volume purchasing.

May 22nd, 2017

The FGC5856-73 branch has generally become known by the first name in the list of equivalents, FGC5856. However on the YFull M222 tree they have picked FGC5862 as the SNP to name the branch and date it currently to 375yo which is probably an underestimate, but by how much is hard to say with any certainty.

https://www.yfull.com/tree/R-FGC5862/

and my sample is kit # YF01405 on that page.

Anyone under M222 and of course FGC4077, FGC5856 and FGC5862 is urged to have their sequencing data analysed by YFull and join the M222 group there

https://www.yfull.com/groups/R-M222 (only YFull customers can see this link).

The growing eminence of the YFull Y chromosome tree has apparently led the designers of the new testing chip at LivingDNA to use the Y3454 name for the FGC4077 branch. Full raw data has yet to be released for early testers of this chip so I will post further information later in the year on this.

The latest on this branch is that FGC5856/62 is still restricted to Perthshire and Angus counties in Scotland based on members who actually have a reliable paper trail within the country. Others who cannot trace their ancestry back to Scotland have surnames so common and widespread that they do not shed any further light on the historical origins of the group. However it is worth emphasising that in terms of tree building and branch identification, a lack of paper trail is of no consequence!

FGC5864 still remains restricted to a sub-group of the Kennedys. It does not appear on any public tree derived wholly from sequencing but is to be seen on my M222 chart at www.kennedydna.com/M222.pdf This mini-branch was found using single SNP data from YSEQ in Berlin.

Technical (sequencing) update – April 2nd 2019

Since my original sequencing test with Full Genomes in 2013 there have been a number of further sequencing tests.

In 2015 I experimented with a low-pass Whole Genome Sequencing (WGS) test again with Full Genomes. At the time this was a cheap route to WGS and held out the promise at least of obtaining reasonably accurate calls for known SNPs. Against this, it has to be considered that no-calls abound and there may be issues with false positives especially for calling new variants. The good news is that out of my original FGC novel SNPs there was only one false negative (FGC5864) and that appears to be caused by a misaligned read. In fact I re-aligned the BAM file to the newer hg38 reference sequence and it still ended up wrong. All the others were either true positives or no-calls. Luckily the market has now evolved and now offers better alternatives for the budget end of the WGS market.

In 2018 I embarked on a series of new test evaluations. The first of these to fully bear fruit was a WGSx30 test from YSEQ, my project's SNP testing lab. This test at an average read depth of 30x on the autosomes (so 15x on the Y chromosome) was conducted at YSEQ's sequencing partner lab in Germany. Given my restricted broadband download facilities I opted to have a hard disk for the full data which contains two 112 Gb FASTQ files, these are the reads pretty much as they emerged from the machine with quality scores but no alignment positions. There is a 32Gb BAM file for the whole genome and smaller extract BAMs for Y and mtDNA. There are also sundry reports including 12 different Variant Call (VCF) files and a spreadsheet of all

the novel SNPs with 'exclusion' reasons for any that were deemed unsuitable. The exclusion codes will be familiar to anyone who has used YSEQ as they match the requisites of the WishASNP product. In particular YSEQ have chosen to call but 'exclude' SNPs in the pseudo-autosomal regions (PAR1 and PAR2). These SNPs may come from a tester's mother's X chromosome. Some companies suppress them and some report them, for example Ancestry test some SNPs in these regions and label them as chromosome 25 in their raw data file. Counting these regions as sequenced Y chromosome can skew comparison statistics where rival companies have chosen to not align or report them. Ultimately it should be up to me to decide how usable the information is.

YSEQ appear to have not found any further novel Y SNPs outside the PAR regions – a testament to how good my 2013 Y Elite test really was!

Finally YSEQ include a composite GEDMatch ready file based on all the major autosomal chip vendors. My file has 1477514 rows of which 633331 are also in my original 23andMe file. This has been uploaded to GEDMatch but flagged as 'research' so others don't have their match list cluttered with duplicates.

I have also had part of my data back from Dante Labs, a new start-up based in the US/Italy. This company is offering some very tempting prices for WGS testing, offset by lengthy delivery timescales. They also initially upset some (not me to be honest) by claiming they would provide download links but later reneging and asking for a fee to get the full BAM/FASTQ data on disk. The VCF reports are provided online and I now have mine, although it is aligned with hg19 despite my advising them some time ago to go straight to hg38. It is possible to convert between the two but it is preferable to await the disk delivery and if the BAM still isn't in hg38 re-align the FASTQ files myself. Many of course want a BAM file to upload to Yfull and placement on the Y chromosome tree; Yfull accept BAM files from a wide range of commercial and scientific sources. This is all the more important now that Yfull have extended tree building activities to include mtDNA which some rival companies deliberately remove from the BAM file in order to sell you a separate product.

Without the actual raw data it is hard to know what to make of the Dante VCF as it has failed to report some of my FGC SNPs but I don't know yet if they were no-calls or just rejected due to quality issues. They have not reported A18822 which was found by Greg Magoon when he re-aligned my Y Elite BAM and they are also silent on two of the FGC58xx SNPs. There is also a separate VCF report for indels which is still to be studied – the YSEQ and Dante tests completed just a few days apart so there is much to look at.

Costs:
YSEQ WGSx30 £1115 including physical disk delivery; turnaround ~ 2 months
Dante WGSx30 299 Euros plus 59 Euros for disk: turnaround ~ 7 months (VCF stage)

There will be a more detailed comparison when I get the Dante disk. The reads are 'only' 100bp but bear in mind that my Y Elite 1.0 test was also 100bp and even when rivals came along later offering 150/151bp reads, none of them have found anything useful that FGC were unable to. Read length tends to be more relevant for longer variants than SNPs of course (indels or STRs).

Since this article is supposed to be about FGC5856 the number of reads obtained for that marker were as follows, in descending order:

Dante WGSx30: 29
FGC Y Elite: 28
YSEQ WGSx30: 14
FGC WGS (low pass): 2

Technical (sequencing) update – May 7th 2019

I have now received the analysis reports from the expensive 10X Genomics Chromium Long Read test from Full Genomes Inc. Briefly this test uses barcoding to tag the molecule each DNA fragment comes from but otherwise sits on top of existing second generation Illumina sequencing. Post sequencing the barcodes can be used to reconstruct ‘linked reads’ much longer than the actual physical 151bp reads. This in turn enables a number of benefits including detection of longer variants which spanned more than one short read, and very importantly it allows the haplotype phasing of the alleles via ‘phase sets’. Some of this information is available in the reports which include a phased_variants.vcf report and further tags can be found in the BAM file.

The benefits of barcoding are not of great application for single base mutations in the non-recombining part

of the Y chromosome which by definition doesn't require phasing (it all came from the father). As there is a tremendous amout of data to analyse my comments so far will be restricted to the Y SNP calling which so far is fairly ordinary. The table above can now be revised thus:

Dante WGSx30: 29
FGC Y Elite: 28
YSEQ WGSx30: 14
10X Genomics: 4
FGC WGS (low pass): 2

Bear in mind that only the Y Elite was a targetted Y chromosome test and that the read depth on the uniparental chromosomes is half that of the autosomes. WGS isn't a test if your top interest is Y chromosome read depth – coverage (width) is another matter and another item that requires expanded comment in due course.

That said, I decided to expand this little comparison to tally counts over all my novel FGC SNPs FGC5856-74 to give a more smoothed out set of statistics. For now I am going to exclude the A18882-6 SNPs that Greg Magoon and I isolated in my hg38 conversion. More on those later too.

Average reads/base over FGC5856-74 (all SNPs except for FGC5874 which is a deletion)

FGC Y Elite: 22
Dante WGSx30: 19
YSEQ WGSx30: 13
10X Genomics: 7
FGC WGS (low pass): < 1

It may be instructive to examine the false negatives ie ancestral reads?

FGC Y Elite: 10%
YSEQ WGSx30: 9%
10X Genomics: 1% (tbc)
Dante WGSx30: 0% (tbc)
FGC WGS (low pass): insufficient data

At the time of writing I lack the BAM files for Dante and 10X both of which report impressively low false negative figures in their respective VCF reports. Of course these will be checked as soon as the BAMs are available.

There is much more to come from the 10X analysis including the STR reports and a full suite of reports from the 10X Genomics Long Ranger pipeline which cover some of the larger scale variation and have revealed a large deletion in my X chromosome…!

https://support.10xgenomics.com/genome-exome/software/pipelines/latest/what-is-long-ranger