Developing computational strategies for assembly of heterozygous DNA sequence data

Student(s)

Su Sarlar
Samuel Lee

Faculty Member(s)

Stuart Lucas

Sabancı Üniversitesi

PURE
PROGRAM FOR UNDERGRADUATE RESEARCH

ABSTRACT

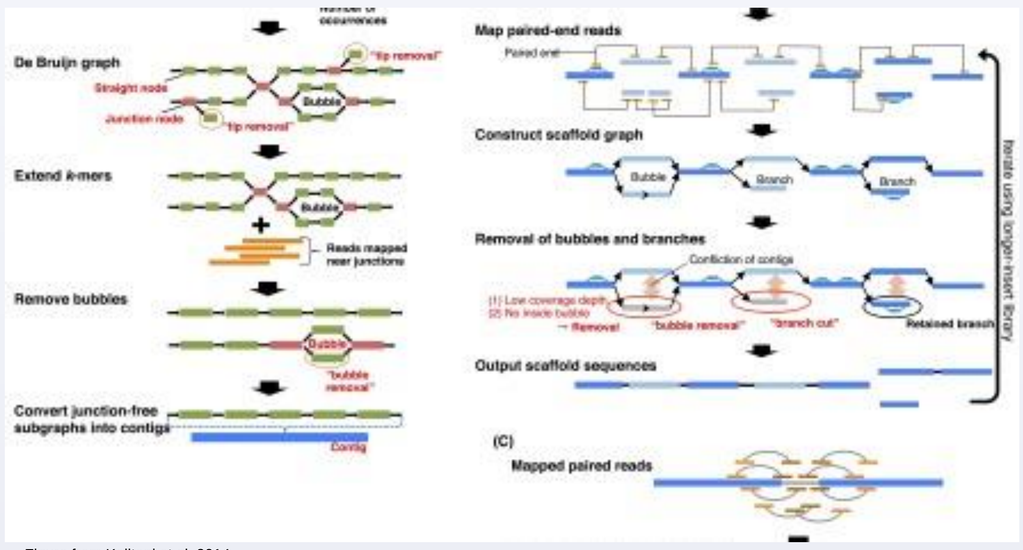


Figure from Kajitani et al, 2014

The process of genome assembly, recently possible due to technological advancements, has presented large amounts of genetic sequence data. However, DNA must first be split into many small fragments which must be read, compared and merged to recover the original genome sequence. Specifically focusing on *Corylus avellana*, also known as the hazelnut, the focus of the project is to work with a large whole-genome sequence dataset, a diploid genome with high heterozygosity, to determine the original genome sequence. Multiple existing programs and software were used to develop new solutions to solve issues of highly heterozygous genomes and create a more integrated and holistic genome assembly.

OBJECTIVES

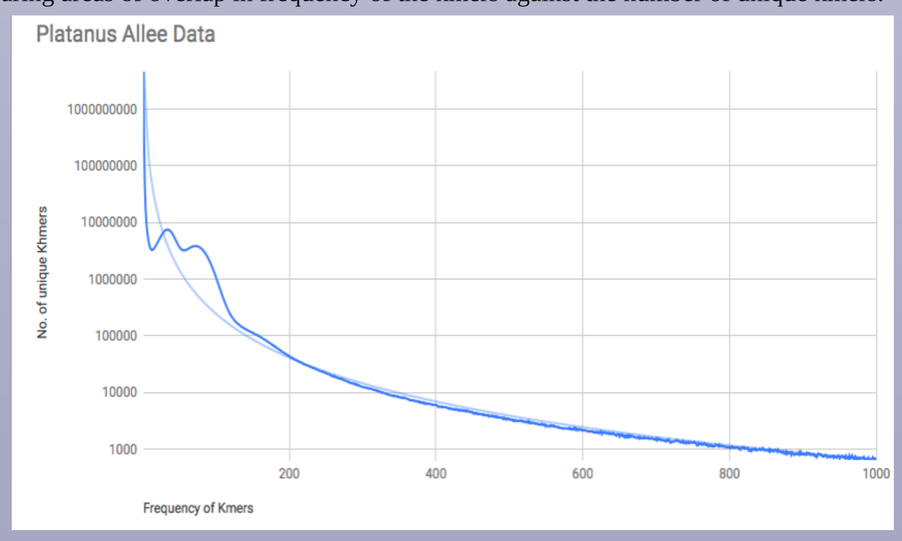
The purpose of this project was to develop a strategy for the assembly of highly heterozygous genomes, specifically working with data of the *Corylus avellana cv.* Tombul. Initial genome assembly of *Corylus avellana*, also known as European hazelnut, produced large numbers of duplicated elements and a larger than expected genome size, implying problems due to heterozygosity. Working on the large whole-genome sequence data and using various existing tools and different software programs, the goal of the project was to solve issues of heterozygosity and develop new ways to filter and present data for a more complete genome assembly.

Through numerous data filtering and analysis procedures, by the end of the project, a data table was created to show the different areas of heterozygosity, with the information of the nucleotide, type of variation (insertion, deletion or single nucleotide polymorphism(SNP)) and starting and ending position of the heterozygous section, each matched with a specific consensus ID.

PROJECT DETAILS

1. Genome Assembly Method In order to assemble our haplotypes, find out homozygous regions where the sequences are likely to be assembled into a continuous string, and heterozygous regions where there are multiple ways that a continuous string can be formed, we have used Platanus-allee, with Nanopore reads and Illumina reads as inputs. The algorithm of Platanus-allee uses De Bruijn graphs to assemble reads into contigs using optimized kmers in this process. Kmers are small words of length k observed more than once in a genomic sequence. Nodes represent kmers and edges represent k-1 overlaps between kmers. Differently than prior assemblers, Platanus automatically extends kmers to handle big and repetitive data. Once contigs are formed, they are scaffolded based on paired end libraries or mate pair libraries. In these contig assembly and scaffolding steps, complicated graph structures are simplified. Contig and scaffold construction is based on graphs without junctions; that is if a node has multiple edges.

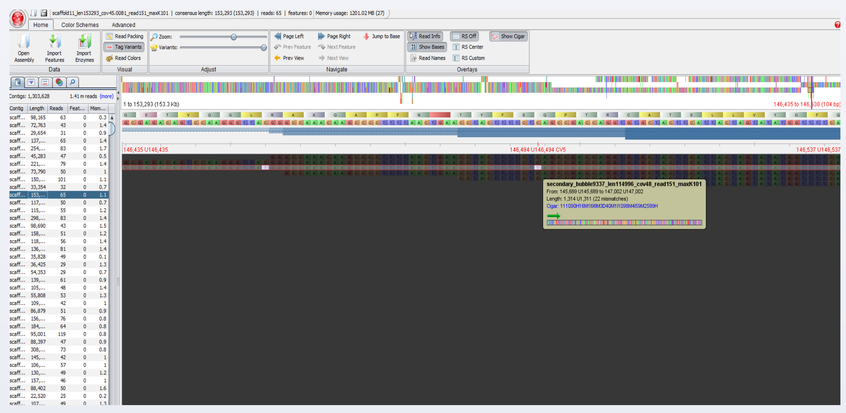
2. Graph of heterozygous areas With the data set obtained from the Platanus-allee, the result was visualized with a graph using a logarithmic scale. X-axis represents the frequency of the kmers and y-axis represents the number of unique kmers. The graph shows areas of heterozygosity in the output by comparing areas of overlap in frequency of the kmers against the number of unique kmers.



3. File Production for genome analysis and assembly The first step included using "bwa- Burrows-Wheeler Alignment Tool": a software package that consists of different algorithms to map low-divergent sequences with a large reference genome (Heng, 2010). This created a .SAM output file, which was further processed and converted to a Binary Alignment/Map (BAM) file using samtools. Samtools uses the Sequence Alignment/Map file format to sort, merge, index and retrieve reads in any region, and is able to import or export files in both SAM and BAM format(Center for Statistical Analysis, 2010). Once the sorted BAM file was created, bcftools – a tool for Binary Call Format (BCF) and VCF – was finally used to create a VCF file.

PROJECT DETAILS

4. Graphical Assistance Throughout this process, another software application called "Tablet" was used to help give further understanding, providing a more visual representation of the sequence alignment map. It revealed the depth of the genome sequence and the areas of overlap and places of potential insertion/deletion to determine the heterozygous parts of the sequence genome (Milne et al, 2013).



FINAL RESULTS

The final table given on the right indicates the starting and ending positions of the heterozygous sites. This table was constructed using 6376 SNVs, each ranking a Phred score of at least 30, so only one in a 1000 SNVs may be an artifact. An additional table was constructed using more than 2600000 SNVs scoring a Phred score more than 20. We have looked at whether the SNVs were much closer than we would expect if they were to be randomly distributed rather than being concentrated in a heterozygous region. If they were particularly concentrated in one place, the beginning and end position of the region would be noted and put in the table on the left.

scaffold name	start pos	end pos
scaffold1057_len154451_cov42.4636_read151_maxK101	133895	136591
scaffold1331_len69105_cov45.0522_read151_maxK101	44465	64515
scaffold14774_len15265_cov44.5163_read151_maxK101	5941	8837
scaffold1509_len131609_cov44.4858_read151_maxK101	27577	68167
scaffold1509_len131609_cov44.4858_read151_maxK101	68167	82042
scaffold156_len124850_cov42.4795_read151_maxK101	30405	31494
scaffold1654_len181247_cov42.8029_read151_maxK101	172252	173314
scaffold1662_len93624_cov41.0913_read151_maxK101	81416	83190
scaffold1843_len91393_cov41.8019_read151_maxK101	54657	59627
scaffold190_len244111_cov46.5638_read151_maxK101	32458	34005
scaffold2185_len158027_cov40.6221_read151_maxK101	50959	118422
scaffold2349_len93518_cov42.4289_read151_maxK101	30180	32645
scaffold2371_len128985_cov45.7464_read151_maxK101	12497	46351
scaffold239_len144546_cov42.1279_read151_maxK101	9486	11892
scaffold2634_len70788_cov40.3694_read151_maxK101	50370	52304
scaffold2724_len55942_cov39.1204_read151_maxK101	14136	17902
scaffold2880_len80250_cov42.4892_read151_maxK101	61668	71702
scaffold3089_len104119_cov41.6663_read151_maxK101	73649	77762
scaffold327_len179894_cov40.9898_read151_maxK101	158624	162018
scaffold3654_len101538_cov43.2634_read151_maxK101	49053	60171
scaffold3654_len101538_cov43.2634_read151_maxK101	60171	62622
scaffold3654_len101538_cov43.2634_read151_maxK101	62622	70277
scaffold3668_len235730_cov41.9994_read151_maxK101	83243	109978
scaffold3668_len235730_cov41.9994_read151_maxK101	109978	181195
scaffold3718_len135558_cov39.2114_read151_maxK101	45415	48837
scaffold3867_len216150_cov41.1996_read151_maxK101	126471	135353
scaffold387_len223330_cov42.0993_read151_maxK101	22613	208275
scaffold3946_len219139_cov40.5865_read151_maxK101	50725	86827
scaffold4633_len103328_cov46.2647_read151_maxK101	87076	88820
scaffold4633_len103328_cov46.2647_read151_maxK101	88820	90092
scaffold4875_len119733_cov41.942_read151_maxK101	108702	109839
scaffold5182_len208670_cov40.5063_read151_maxK101	8953	97629
scaffold5182_len208670_cov40.5063_read151_maxK101	97629	129096
scaffold5283_len67208_cov38.3336_read151_maxK101	43124	46174
scaffold6038_len178800_cov40.9859_read151_maxK101	8366	40802
scaffold6038_len178800_cov40.9859_read151_maxK101	40802	72546
scaffold6038_len178800_cov40.9859_read151_maxK101	72546	144663
scaffold6060_len42394_cov37.4162_read151_maxK101	3325	5323
scaffold6156_len155746_cov40.9537_read151_maxK101	53712	97595
scaffold637_len75290_cov45.086_read151_maxK101	26339	52755
scaffold6528_len49544_cov44.8359_read151_maxK101	12004	13617
scaffold6528_len49544_cov44.8359_read151_maxK101	13617	15620
scaffold6837_len115042_cov40.9749_read151_maxK101	24333	78709
scaffold6950_len143390_cov43.8628_read151_maxK101	41132	65973
scaffold6950_len143390_cov43.8628_read151_maxK101	65973	69498
scaffold7027_len57151_cov47.5594_read151_maxK101	29975	31220
scaffold727_len187555_cov41.1165_read151_maxK101	22255	112753
scaffold727_len187555_cov41.1165_read151_maxK101	112753	170767
scaffold7425_len100831_cov45.8154_read151_maxK101	68682	71528
scaffold7595_len90015_cov41.1146_read151_maxK101	81173	88018
scaffold7603_len48857_cov53.7836_read151_maxK101	19578	20981
scaffold864_len114662_cov40.984_read151_maxK101	31961	98906
scaffold87_len49417_cov39.5279_read151_maxK101	33107	34387
scaffold879_len236274_cov45.6087_read151_maxK101	65315	116177
scaffold913312_len20233_cov44.1781_read151_maxK101	13952	14976
scaffold913363_len7309_cov56.4251_read151_maxK101	131	4828
scaffold913363_len7309_cov56.4251_read151_maxK101	4828	5923
scaffold929_len294974_cov45.1665_read151_maxK101	52762	182619
scaffold9645_len74061_cov39.2884_read151_maxK101	1647	35119
scaffold997_len252593_cov43.8599_read151_maxK101	126491	128448

CONCLUSIONS

By the end of the project, we were able to successfully identify the heterozygous regions in the genome. The dataset we have used included SNP's with a Phred score above 30: only one in every 1000 reported SNVs would be an error. However a large proportion of the SNVs given the final vcf data ranked a Phred score of about 25, indicating that about one in 316 SNV reportings may actually be an error. We also constructed the table using the data of SNVs with a Phred score of at least 21; indicating one in 125 reportings may be an error. 2673571 SNVs was used for this table. The latter table may be more throughout and dependable.

For further improvements, there are upcoming and developing softwares: such as ones promising to work with lower coverage, but unable to handle repetitive regions yet, or previously released softwares being improved such as Meraculous-2D (Goltsman 2017). Based on the current literature, we think that haplotype sensitive genome assemblers are quickly developing and improving, and genome assemblies in the future will be much easier and dependable.

Our results were necessary for proceeding to next steps of genome assembly of Tombul cultivar of the hazelnut. Currently, we have yet to find out what haplotypes have got which heterozygous regions.

In conclusion, using specific conditions during the filtering process, including depth and quality, the first final data table obtained was smaller than expected, which allows for further research and testing in different filtering processes and conditions to acquire a more realistic genome sequence.

REFERENCES

National Human Genome Research Institute. An Overview of Human Genome Project. (2016, May 11). Retrieved from ttps://www.genome.gov/12011238/an-overview-of-the-human-__genome-project/.

Center for Statistical Analysis. (2013, 26 February). BAM. Retrieved from https://genome.sph.umich.edu/wiki/BAM

Garg, S., Rautiainen, M., Novak, A.M., Garrison, E., Durbin, R., & Marschall, T. (2018). A graph-based approach to diploid genome assembly. Bioinformatics. Volume 34, Issue 13, 1 July 2018, Pages i105–i114, https://doi.org/10.1093/bioinformatics/bty279

Goltsman, E., Ho, I.Y., & Rokhsar, D. (2017). Meraculous-2D: Haplotype-sensitive Assembly of Highly Heterozygous genomes.

Harismendy, O., Ng, P., Strausberg, R., Wang, X., Stockwell, T., Beeson, K., Schork, N., Murray, S., Topol, E., Levy, S., Frazer, K. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome biol. 2009;10:R32

Heng, L. (2011, July 5). Manual Reference Pages - samtools (1). Retrieved from http://samtools.sourceforge.net

Heng, L. (2010, February 28). Burrows-Wheeler Aligner. Retrieved from http://bio-bwa.sourceforge.net

Heng, L. Bcftools. Retrieved from http://www.htslib.org/doc/bcftools.html

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, Kohara Y, Fujiyama A, Hayashi T, Itoh T, "Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads" Genome Res. 2014 Aug;24(8):1384-95. doi: 10.1101/gr.170720.113.

Li & Durbin 2009, Bioinformatics 14:1754-60).

Milne, I., Stephen, G., Bayer, M., Cock, P., Pritchard, L. Cardle, L., Shaw, P., Marshall, D. (2013). Using Tablet for visual exploration of second-generation sequencing data, Briefings in Bioinformatics, Volume 14(2), pp.193-202.