Error Correction and allele/SNP calling for the whole genome assembly of *Spironucleus salmonicida*

Mayank Mahajan
Institute of Cell and Molecular Biology, Uppsala University, Sweden

Abstract: *Spironucleus salmonicida* is a unicellular eukaryote with two nuclei and it is a diplomonad. The main goals of this project were to remove errors from the assembled genome and to detect alleles/SNPs in the genome. The genome had already been assembled using the single and mate pair reads from '454 Titanium'. The same reads that were used for *de novo* assembly were mapped back on the genome. Illumina paired reads were also mapped in addition for the indel detection. The variance in the mapped bases at each position was used to detect SNPs/Alleles and error correction. The most critical decision is to classify the variance as an error or an SNP. The results showed considerable improvement in the quality of genome. The number of alleles and SNPs detected were quite close to the expected number.

Background and Introduction

Spironucleus salmonicida is a species of fish parasite. These parasites create foul-smelling, puss-filled abscesses in muscles and internal organs of the farmed salmon. Our research group is working on the whole genome of this species. This genome project could be useful in various aspects. For example, to locate the genes responsible for its infectious behaviour and design the drugs to tackle the problem that salmon farms are facing right now. The two main goals of this project were to remove errors from the assembled genome and to detect alleles/SNPs in the genome.

Spironucleus salmonicida genome is thought to be about 12 Mbp long. It is a eukaryote and a diplomonad which means it's a unicellular organism with two nuclei. Its genome contains the minimalistic set of genes that are required for the basic functions of a single organism [2]. In previous studies very less allelic variation has been found within genomes of similar species[1]. This information is quite important to infer the sexual or asexual behaviour of the species. One of the goals of this research is to show allelic variation in *S. salmonicida* genome. It's not yet established whether there is any recombination between the chromosomes in 2 different nuclei. In this study it is assumed that the species may have up to 4 alleles or alternate bases at every single position.

The *S. salmonicida* genome was assembled *de novo* using mate pair and single ended reads from 454 sequencing. A 454 Titanium device was used for the sequencing and Celera assembler was used for the assembly.

As we know the problem with 454 machines is that it gives indel errors. The aim was to remove indel errors from the assembled genome by using Illumina reads. Illumina reads, on the other hand, have more base calling errors. Hence, a hybrid approach was needed to get a high quality sequence and make full use of the variety of data we have. The approach followed in this article uses both the 454 and Illumina reads together. Illumina reads are preferred to detect indels and 454 reads get preference for detecting the base calling errors.

The allele/SNP calling was done simultaneously with the error removal process. Further in the article we shall see that how both the processes go hand in hand. Actually, Allele/SNP calling would be the first step and then we shall try to do an optimal filtering to minimize the false positives and true negatives. In the mean while it should be made sure that all the other important information in data is not lost.

Materials

The reference genome available in the beginning was assembled *de novo* using mate pair and single end 454 reads. The assembly was performed using Celera Assembler. The reference has 232 Scaffolds and 12,885,212 bp. N50 for the assembly is 151,388 bases. The 454 reads used for assembly are 80.18%. The Average Coverage for Assembly is 47.73 reads per position.

In the beginning single and mate pair reads from '454 Titanium' were available for analysis (in sff format). These reads are 500 base pairs in length with insert size of 3 kb in case of mate pairs. These are the same reads used for the *de novo* assembly. And the Illumina pair end reads were available in fastq format. Average length was 100 bp with 350 bp inserts.

Model and Methods

The whole process that was followed in this project can be explained in nine main steps. The project was finished within the two and a half months, and therefore, most of the steps/methods in the Model were indentified while working on the whole process. Some of the steps were identified quite late and could not be included in our current analysis. All these steps will be implemented in the future projects. *Table 1* gives main methods used for the whole process. A detailed description of those steps follows

Step 1: Extract and convert all reads to fastq format which is supported by *bwa* and *ssaha2*. All the reads from the 454 device were in the sff format. First they were converted into fastq format. 'sff_extract' utility was used

for this purpose. Next mate pair reads from the 454 device were supposed to be separated because the alignment software required the input in that way. Biopython [4] library was used to develop scripts to separate the forward and reverse mate pair reads from 454 and sort them [Supplement 1]. The Illumina reads were already in fastq format.

Table 1: This shows a concise overview of the overall process, presented in a hierarchical order.

Main Steps in the process	Implemented in Project
1. Extract and convert all reads to fastq files	Yes
2. Normalize the quality values in different kinds of Sequencing data	Future Task
3. Mapping the reads. Paired/single	Yes
4. Merge all mapped reads from similar sequencing technologies	Yes
5. Pileup using deprecated Samtools utility or Write a script using some bam parser.	Done using Samtools
6. Get aggregate data from individual pileup files (no. of bases, supporting reads, indels, mean quality etc.)	Yes
7. Filter the data below certain quality level. Use statistical methods.	Yes
8. Do some final filtering in and around indels with high support.	Partially Implemented
9. Generate new consensus based on data and show the alleles/SNPs	Yes

Step 2: The quality values from different sequencing technologies are ranged differently. In our case Illumina reads had quality values ranging from ASCII 64 to 105. The 454 reads had quality values varying from ASCII 33 to 74. The alignment tools assume the quality value to be in Sanger format, i.e. quality values ranging from ASCII 33 to 128. Hence, the quality values from Illumina reads should be reduced by 31. As the range of quality from both technologies is 40 we don't need to scale them. But, these adjustments were not done for our project as it was realised quite late. These things don't have much overall effect except that we are forced to adopt double standards later for Illumina and 454 reads.

Step 3: Illumina paired reads we aligned to the available reference genome using *bwa* [5] tool. The reads below an average quality value 15 were filtered. Paired reads were mapped to the reference using the maximum gap of 450 bp. Then the aligned reads were sorted using *Samtools* [3]. During PCR amplification some regions get duplicated too many times. These duplicate reads must be removed because these are useless and do not give use any useful information. *Samtools* was used to remove these duplicates. 91.4 percent of paired reads from Illumina were aligned and the mean coverage was 239.7 reads per position.

Similar process was used for the 454 mate pair and the single end reads. However, a different tool, *ssaha2* [6], was used for mapping 454 reads because it can handle the longer read lengths and is a preferred tool for mapping 454 reads. *bwa* is only recommended for Illumina reads. Insert size for the mate pair reads was 3kb with a standard deviation of 900 base pairs. The mapped reads were sorted and duplicates were removed from among them. 73.7 percent of single reads and 79.1 percent of mate pair reads from 454 were aligned the mean coverage was 46.3 reads per position. Remove duplicate reads that are generated due to amplification during the PCR reactions.

Step 4: The bam files generated for 454 mate pair and single reads were merged together using Samtools. Samtools were used again to remove duplicate reads from mapping data.

Step 5: The deprecated pileup option in *Samtools* was used on both the data sets (Illumina paired end, 454 combined data).

Step 6: Then important information was extracted. The same includes complete indel info, coverage, number of alternate bases, number of reads supporting the alternate bases and the average quality value for each alternate base. It was quite hard to implement as the pileup format is not well documented. This was done using programs written in C [Supplement 2 – source code, input, and output]. Later, it was realized that the same would have been easily done using some sam/bam parser API (e.g. sam.h/bam.h). Some initial filtering was also done in this script. The likelihood of indel should be greater than 0.8 times the likelihood of reference being correct and the indels must have at least two reads to support them. All the indels not satisfying these criteria were removed. The likelihood values mentioned above were calculated by Samtools during the pileup.

Step 7: Another C Program was used to combine both the data sets and filter the data [Supplement 3 – source code, input, and output]. But before filtering it was important to decide that how much importance should be given to the quality of reads and how to use that information. We had mean quality values for each base at a position. We want to use this information such that the bases with high mean quality should get a higher preference. The quality value for each base was scaled to a range [1, x] and then multiplied to the number of reads supporting the base. If a base has maximum quality then the number of supporting reads for that base will be multiplied by 'x'. If a base has minimum quality then the number of supporting reads for that base will be multiplied by 1 and so on. The value we choose for 'x' will decide how much importance the quality will get. A higher value of 'x' gives more

significance to the quality values. In this project the value of 'x' was chosen to be 3. This gives us a new attribute in which scaled quality values have been multiplied by the number of supporting reads. The same will be referred to as 'cq' from now on. Unfortunately, we did not have reliable quality values for the indels, hence, 'cq' in their case just means the number of supporting reads for that indel.

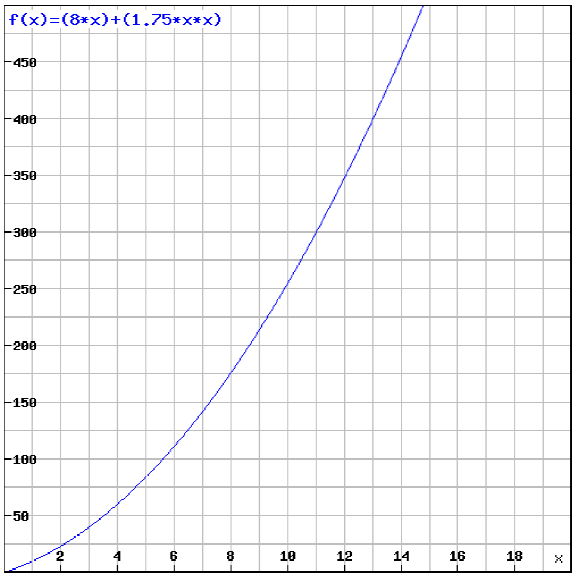

Figure 1: The cut-off curve for alternate bases. E.g. If coverage is 100 at a position then the value of 'cq' must be at least 6 for an alternate base. Which means coverage (100) at this position should be less than or equal to $8cq+1.75cq^2$.

The bases and indels at each position were then filtered according to the 'cq' value attached to them. A very basic filtering would involve excluding bases with 'cq' less than a chosen constant value or a certain proportion of coverage at each position. Optimally, a different cut-off value might be required for each position depending on the coverage at the position. An approximate curve was needed which could represent *the total coverage at a position* plot against *the cut-off values of 'cq'*. A quadratic equation was a good approximation of such a curve. *Figure 1* and *Figure 2* show the curves used for filtering the alternate bases and indels respectively. At higher values of coverage the fraction of support required for an indel or alternate base to get selected is less. For example, in coverage of 10 you need 2 supporting reads to keep an alternate base (which is 20 percent), whereas, with coverage of 200, 15 supporting reads are enough (which is only 7.5 percent). As we were going to filter the bases/indels on the basis of the new attribute 'cq', the equation $\{a*cq +b*cq^2\}$ was used. The values of 'a' and 'b' can be varied and tested at various level. The number of false positives and false negatives we get would be determined by our choice of parameters. Figure 1 and figure 2 show the values of 'a' and 'b' that we used respectively for bases and indels, after much thinking and discussion. In this step we have tried to retain the false positives here so that we do not lose any false negatives. Hence, the cut-off filters are a bit lenient.

At last some of the probability and likelihood values were calculated for each alternate base from each dataset and stored in the output file. Probability, that a base is the only base at this position ($P_{Illumina}$, P_{454}).

$P = cq / \sum cq$; $P_{avg} = (P_{Illumina} + P_{454})/2$

Though, it is the probability of base being the only base at the position, we will use this value for calling SNPs in the next Steps.

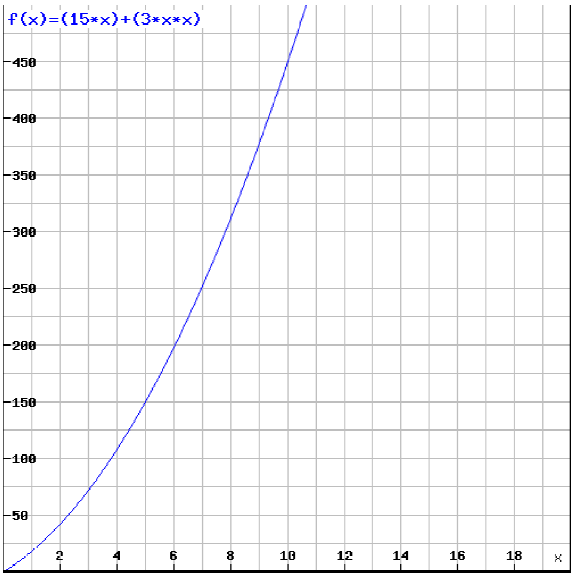

Figure 2: The cut-off curve for indels is steeper which means that less supporting reads are required. Y – Axis shows the total coverage at this position. X – Axis shows the minimum value of cq (same a no. of supporting reads in case of indels) required for an indel to be retained.

The cut-off function was modified a bit to get so likelihood score of base being an alternate base ($L_{Illumina}$, L_{454}).

$L = (a*cq +b*cq^2)/coverage^{(1+E)}$

$L_{Total} = L_{Illumina} + L_{454}$

As we know cut-off function should give a value greater than coverage at that position. Hence, 'L' should be greater than or equal to 1 in that case. This equation is more reliable than the probability as it uses coverage and grows in a quadratic manner with cq. But, when the coverage goes too high it might be build some false confidence which we don't want. Hence, the coverage is raised to the power of a small fraction which reduces the value of 'L' exponentially as coverage goes too high. In this case 0.01 was chosen as the value of E.

All the bases with likelihood less than 1.0 were removed from the datasets. At positions where both Illumina and 454 reads were mapped, sometimes, each dataset (Illumina/454) shows a different set of alternate

bases. I such a case only the alternate base present in both the datasets was retained and the rest were removed. The average of the likelihood scores from both datasets was used for such alternate bases. Both the dataset were handled separately during this whole step and were not merged. The output, however, was in a merged format.

There is a flaw in the likelihood values 'L'. The coverage of Illumina dataset is around 6-7 times the coverage of 454 dataset. This can be seen in the likelihood values where the $L_{Illumina}$ is consistently around 6-7 times L_{454}. This will be improved in future projects. In this project the Probability values were used in the further analysis and filtering.

Step 8: For future projects it is suggested to do stricter filtering in the region neighbouring an insert or a deletion as these regions have higher mapping errors. The size of the neighbourhood is proportional to the size to the indel. In this project only the region right next to the insertions was filtered more strictly. The details are given in step 9.

Table 2: An ordered list of the main functions performed at this step and issues tackled.

1. If a position is unmapped use the base in the old reference genome
2. Rank all the bases at a position using Probability values 'P' from last step
3. The base at rank one should go into new reference. But, if the position is not mapped by both datasets then we use the base in old reference
4. If no bases are selected then write 'N' to the new reference.
5. If more than one base is present at a position then write them to alleles file
6. Ignore an indel in 454 dataset if Illumina reads are mapped at the position.
7. Rank all the indels at a position using Probability values 'P' from last step
8. The indel at rank one should go into new reference if its 'P' value is more than that of old consensus base.
9. If an insertion of length L take place then ignore all indels, and bases with P less than C, in next L reads
10. If an indel passes through all the filtering then write it to the alleles file.

Step 9: Last step was to do the final filtering of data to get the final list of alleles/SNPs and to generate a new corrected consensus. *Table 2* shows the main functions of this script in a hierarchical order. The table also explains all the crucial decisions taken in during SNP/allele calling and generate a corrected reference genome. Illumina reads were preferred for decision regarding indels. As the output of this step some statistics were generated with the main allele file and a new reference file. These functions were performed using another C Program [Supplement 4 provides the source code of the C program that performs this step and the details about the inputs and outputs]

Results

A lot of analysis can be done using the information available in the final four output files. The following information was generated by doing some quantitative analysis on files. There were 232 total scaffolds in the old reference and their number remains the same in the new reference as expected. We were only mapping the reads on the old reference which cannot change the number of scaffolds. The old reference was 12,946,900 bp long, whereas, the new reference has 12,948,667 bp. The genome is now longer by 1,767 bases. It suggests that there have been more insertions than deletions in the genome correction process. The statistics generated by the program should be able to authenticate these numbers.

Table 3, 4 and *5* show some important statistics computed by using the information in the output files.

Table 3: The statistics generated for the insertions, or deletions, which were added to, or subtracted from, the old reference.

Indels	3,845
Indels from 454 dataset only	0
Indels from Illumina/Both datasets	3,845
Inserts	2,895
Insert length	3,360
Deletions	950
Deletion length	1,593
Total 1 bp insertions	2710
Total 2 bp insertions	100
> 2 bp insertions	85
Total 1 bp deletions	757
Total 2 bp deletions	104
> 2 bp deletions	89

The total insertion length is 3,360 and the deletion length is 1,593. Therefore, the genome should be 1,767 base pair longer. This authenticates our previous calculation of the genome length. Another important fact to be observed is that due to the stricter filtering in step 8(6^{th} row in Table 2) and low probability values of 454 indels none of them were incorporated into the new reference. The reference was not altered at the positions with only one dataset mapped.

Table 4: Statistics generated for the positions where the reference was altered.

Total positions	1,763
Positions with only 454 mapping	0
Positions with only Illumina mapping	0
Positions with both mapping	1,763

Overall, there were 23973 positions in the whole genome with only 454 mapping. And, there were 735 positions in the genome with only Illumina mapping. 12,864,748 positions in the reference genome were mapped by both the datasets (Illumina and 454). There were 59,044 positions where none of the Illumina/454 reads were mapped. For all such cases, the consensus base from the old reference was used in the new reference.

There were 167 positions where no definite base could be chosen due to very less support to each of the bases. The consensus base from the old reference was used for all such positions. I would soon do some more analysis on the final results and generate more information which is still missing in this data.

Table 5: No. of alleles/SNPs called by the algorithm:-

Total	54,325
Indels	4,450
Indels in 454 dataset only	594
Inserts in 454 dataset only	312
Deletions in 454 dataset only	282
Indels in Illumina/both dataset	3,856
Inserts in Illumina/both dataset	2,903
Deletions in ill/both dataset	953
SNPs	49,875
Positions with only Illumina mapping	538
Positions with only 454 mapping	2,898
Positions with both mapping	46,439
Positions with 1 alternate (2) bases	47,976
Positions with 2 alternate (3) bases	1,840
Positions with 3 alternate (all 4) bases	58

Conclusions and future work

It was observed throughout the genome that wherever reads from Illumina were not mapped the number of indels were high. And wherever the reads from 454 were not mapped the number of alternate bases was high. It would be really informative to generate some graphs where these things could be seen appear more clearly. Some graphs can be drawn from this data showing the variations in error detection, indels and SNP/allele calling due to variation in coverage from Illumina and 454 reads. Various other graphs can also be generated (example: error frequencies in different regions, difference in results by varying the parameters that we discussed). A lot of parameters have been used in this project in various equations. It will be really advantageous to try different values of these parameters and check how the quality of results differs. E.g. Change sensitivity/specificity of SNP/allele calling by using different parameters for the cut-off curve. It would informative to generate an extra statistic which specifies if the coverage from one dataset (454/Illumina) is too less. Though, the same data can be generated from the output files that we get after the final step.

Acknowledgements

I am thankful to Jan Andersson and Feifei Xu for their guidance and support. I am also thankful to Jan Andersson for giving me the opportunity to work on this project.

References

1. Andersson JO, Sjögren ÅM, Horner DS, Murphy CA, Dyal PL, Svärd SG, Logsdon JM Jr, Ragan MA, Hirt RP, Roger AJ. A genomic survey of the fish parasite *Spironucleus salmonicida* indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. *BMC Genomics* 2007, 8: 51. doi: 10.1186/1471-2164-8-51

2. Roxström-Lindquist K, Jerlström-Hultqvist J, Jørgensen A, Troell K, Svärd SG, Andersson JO. Large genomic differences between the morphologically indistinguishable diplomonads *Spironucleus barkhanus* and *Spironucleus salmonicida*. *BMC Genomics* 2010, 11:258. doi: 10.1186/1471-2164-11-258

3. Samtools - http://samtools.sourceforge.net/ - Date visited 28 August 2011

4. Biopython - http://biopython.org/wiki/Main_Page - Date visited 28 August 2011

5. Burrows-Wheeler Aligner - http://bio-bwa.sourceforge.net/ - Date visited 28 August 2011

6. ssaha2 - http://www.sanger.ac.uk/resources/software/ssaha2/ - Date visited 28 August 2011