A Modified Genovo Metagenome Assembler for 454 Paired End Reads

Afiahayati[1,a)] Tsuyoshi Hachiya[1] Kengo Sato[1] Yasubumi Sakakibara[1,b)]

Abstract: Metagenomes have presented assembly challenges, how to assembly of multiple genomes from mixed sequence read of multiple species. Single genome assembler is not sensitive enough when applied in this case. Genovo is a metagenome assembler under a generative probabilistic model which covers more bases and recovers more genes than the other methods, but it is designed for 454 single read. Paired end sequencing is currently widely-used in metagenome project. In this research, we attempt to modify Genovo for paired end read utilizing mate pair information. First, we extend to add bonus parameter in chinese restaurant process used in Genovo to get a prior accounts for the unknown number of genomes in sample. This bonus parameter intends a pair of reads should be in a contig so that it can be one of the efforts to solve chimera contig case. Second, for sampling process of a read location in a contig, we use relative distance of a read to its mate considering the insert length instead of using distance between offset and center of contig which is used in Genovo. Using this related distance, a read will be mapped in correct location.
We demonstrate the performance of our modified Genovo by comparing it with the original Genovo itself. We use simulated metagenomic dataset of 13 virus generated by Metasim with different community complexity. Our strategies can work well achieving better performance for paired end read and the computational cost doesnt increase, same with Genovo.

Keywords: metagenome, assembly, genovo, 454, paired end read

1. Introduction

Next generation sequencing (NGS) technologies have allowed an explosion in sequencing with the increased throughtput and decrease in cost of sequencing [1]. The field of metagenomic has adapted to the new type of sequencing technologies which allows us to generate reads from multiple genomes effectively [2]. While a number of metagenomes have been sequencing using NGS technologies, only few works succeeded reporting their assembly result [3],[4],[5]. Metagenomes have presented a number of additional assembly challenges, how to assemble multiple genomes from mixed sequence read of multiple species. The challenges are from uncertainty about the populations size and composition [6]. Multiple genomes are represented disproportionately owing to uneven community composition resulting in poor or no coverage of many parts of many genomes [1]. Single- genome assembler is not sensitive enough when applied in this case.

There are a number of effective assemblers for single genomes, but only five (MetaVelvet, MetaIDBA, Genovo, MAP and IDBA-UD) attempt to solve metegenome cases. Metavelvet, MetaIDBA and IDBA-UD use De Bruijn graph approach. They was designed to handle short read data. MetaVelvet attempts to bin genomes using graph connectivity and coverage(abundance) difference [5]. MetaIDBA bins genomes based on an important observation, then for each bin, captures the slight variants of the genomes of subspecies from the same species by multiple alignments and represents the genome of one species, using a consensus sequence [2]. IDBA-UD is an extended of MetaIDBA solving the problem that sequencing depth of different regions of genomes from different species are highly uneven [7]. MAP uses an improved OLC(Overlap/Layout/Consensus) strategy and mate pair information. MAP was designed for reads by Sanger and 454 sequencing [8]. Genovo is a metagenome assembler under a generative probabilistic model. It performs a series of iterated deterministic and stochastic hill climbing moves, based on the iterated conditional modes (ICM) algorithm. Different with the other methods, Genovo does not throw away any reads hence is able to extract more information from the data, which contributes to the discovery of more low-abundance sequences. Genovo is designed for 454 sequencing data [6].

The sequencing technologies producing longer reads, such as 454 sequencing (usually 200-500 bp), are still the overwhelming recommendation and thus remain the major source of metagenomic sequence data [8], [9]. Paired end sequencing is currently widely-used in metagenome project. Genovo covers more bases and recovers more genes than the other methods, even for low-abundance sequences, but it is designed for 454 single read. Therefore, we attempt to modify Genovo for paired end read utilizing mate pair information.

2. Overview of Genovo

Genovo is a metagenome assembler under a generative probabilistic model. An assembly is represented as a list of contigs and

[1] Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan
[a)] afia@dna.bio.keio.ac.jp
[b)] yasu@bio.keio.ac.jp

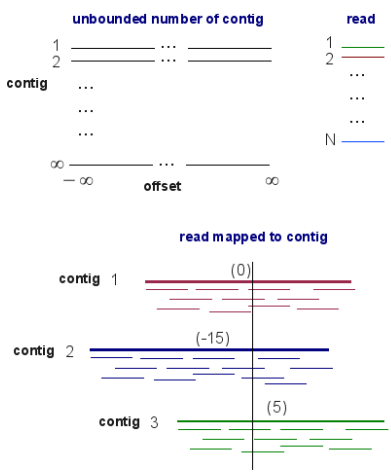

Fig. 1 The Generative Probabilistic Model in Genovo

a mapping of each read to contigous area in a contig. Each contig is represented as a list of DNA letters $\{b_{so}\}$, where b_{so} is the letter at position o of contig s. Each read x_i has its contig number s_i and its starting location o_i within the contig. The alignment(orientation, insertions and deletions) required to match x_i base-for-base with the contig is denoted by y_i. Bold-face letters, such as **b** or **s**, represent the set of variables of that type. The generative probabilistic model is illustrated in **Fig. 1**. There are N reads mapped to 3 contigs. The probabilistic model is described as below

(1) Construct an unbounded number of contigs (each has unbounded length)
Assuming that there are infinitely many of contigs. The number of read is finite, so only a finite number of infinitely many contigs will have any reads assigned to them.

(2) Assign place holders for the beginning of reads in a coordinate system of contigs and offsets
There is a coordinate system of contigs and offset showing the position of reads mapped in contigs. There are two steps to map the reads in the contigs, first, partition the reads to cluster and then assign each cluster of reads to a contig. A chinese restaurant process (CRP) is used to generate the clusters as a prior accounts for the unknown number of genomes in the sample, shown in formula 1.

$$s \sim CRP(\alpha, N) \tag{1}$$

The contigs are treated as infinite in length, from minus infinity to infinity. A good contig is defined as a contig having the most reads toward the center of contig. Therefore, a starting point of read oi within each contig is assigned using a symmetric geometric distribution, shown in formula 2 .

$$o_i \sim G(\rho_s) \quad \forall i = 1..N \tag{2}$$

The detail of CRP and the symmetric geometric distribution will be explained in the Methods section.

(3) Copy each reads letters (with some noise) from the place it is mapped to in the contig
The read letters x_i are copied(with some mismatches) to the contig starting from position o_i and according to the alignment y_i (encoding orientation, insertion and deletion), shown in formula 3, l_i is the length of $read_i$, ρ_{ins} is the probability of insertion, ρ_{del} is the probability of deletion and ρ_{mis} is the probability of copied incorrectly(mismatch).

$$x_i, y_i \sim A(l_i, s_i, o_i, \mathbf{b}, \rho_{ins}, \rho_{del}, \rho_{mis}) \quad \forall i = 1..N \tag{3}$$

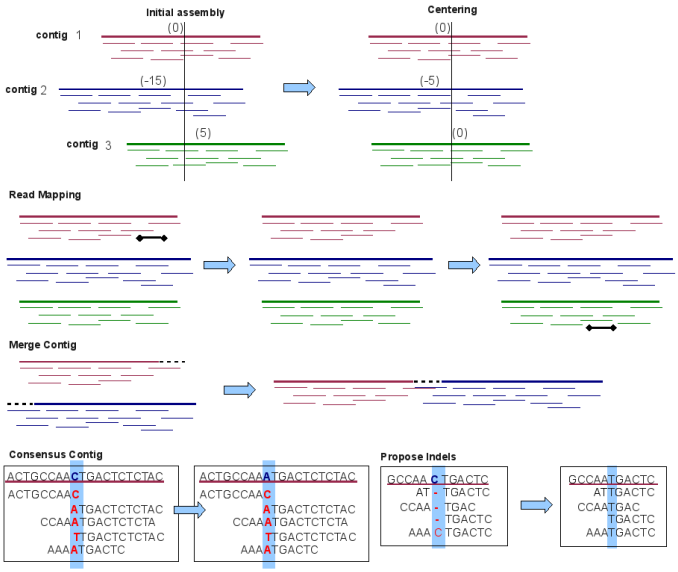

Fig. 2 Iterative Procedures of Genovo

To discover appropriate assemblies, Genovo performs a series of deterministic and stochastic hill-climbing moves in an iterative fashion to reach the best likelihood. The likelihood of this model consists of the likelihood of the alignments $logp(\mathbf{x,y|s,o,b})$, the likelihood for generating (uniformly) each contig letter $logp(\mathbf{b})$, the likelihood of contigs $logp(\mathbf{s})$, and the likelihood of offsets $logp(\mathbf{o|s},\rho)$, shown in formula 4,5,6,7 and 8

$$logp(\mathbf{x,y|s,o,b}) + logp(\mathbf{b}) + logp(\mathbf{s}) + logp(\mathbf{o|s},\rho) \tag{4}$$

$$logp(\mathbf{x,y|s,o,b}) = \sum_{i=1} score^i_{READ} \tag{5}$$

$$logp(\mathbf{b}) = -log|\beta|L \tag{6}$$

$$logp(\mathbf{s}) = S\,log(\alpha) + \sum_{i=1}^{S} log\Gamma(N_s) + const(\alpha, N) \tag{7}$$

$$logp(\mathbf{o|s},\rho_s) = \sum_{i=1}^{S}[O_s log(1-\rho_s) + N_s log\rho_s + const(N)] \tag{8}$$

where S is the number of contigs, N_s is the number of read in contig s, $O_s = \sum_{k=1}^{N_s} |o_k|$, L is the total length of all contigs, ρ_s is the control parameter of the length of a contig, β is the count of DNA character =4 and $score^i_{READ}$ is the alignment score or $read_i$ mapped to the contig.

The moves, can be said the procedures, are based on the iterated conditional modes (ICM) algorithm maximizing local conditional probabilities sequentially, in order to reach the MAP solution. The algorithm is run until convergence (200- 300 iterations). Genovo outputs the assembly that achieved the highest probability thus far. The illustration is shown in **Fig. 2**. The procedures

are described as below

(1) Consensus contig
This procedure performs ICM updates over the (observed) letter variable b_{so} by setting the contig to be the consensus contig of the reads in their current mapping. This procedure aims to increase the likelihood of alignment.

(2) Read mapping
This procedure performs stochastic ICM updates over the read variables s_i, o_i, y_i. It is the main procedure in Genovo. A new location of a read (s_i, o_i, y_i) is choosen by sampling from the joint posterior $p(s_i = s, o_i = o, y_i = y | x_i, \mathbf{y}_{-i}, \mathbf{s}_{-i}, \mathbf{o}_{-i}, \mathbf{b}, \rho)$. In the illustration, a read from the contig 1 moves to the contig 3.

(3) Global moving
These procedures speed up convergence which changes a set of variables at once.
(a) Propose indels
If at a specific location most reads have an insertion then propose to delete the corresponding letter in the contig and realign the reads. The proposal will be accepted if improving the likelihood.
(b) Center
Each contig has a center. A good contig is defined as a contig having center towards zero. This procedure is shifting the coordinate system of each contig to maximize the $p(o)$ component of the likelihood, making the center of contigs towards zero. In the illustration, there are 3 contigs. After implementing this procedure, the center of each contig is shifting toward zero.
(c) Merge
Merging two contigs whose ends overlap (20 nucleotides), if it improves the likelihood.

(4) Chimeric solving
Chimeric reads are reads having two segments of length ¿ 20 that mapped to noncontiguous portions of the reference genome [10]. Inclusion of these in the assembly may create chimeric contigs which spuriously join segments from two different genes [11]. Genovo algorithm assumes that these reads often find their way to the edge of an assembled contig. In order to solve this case, Genovo occasionally (every 5 iterations) disassembles the reads sitting in the edge of a contig in order to allow other correct reads or contigs to merge with it so that the likelihood will increase.

3. Methods

We attempt to modify some procedures of Genovo in order to fit in with 454 paired end read utilizing mate pair information. First, we modify CRP by adding a bonus parameter. This bonus parameter intends a pair of reads should be in a contig and also as one of the efforts to solve chimera contig case. Second, we modify the sampling process used in the read mapping procedure. We use relative distance of a read to its mate considering the insert length instead of using distance between offset and center in sampling process of a read location in a contig. The detail methods are explained in the next sections below. We dont use a chimera solving procedure from Genovo. In our model, more

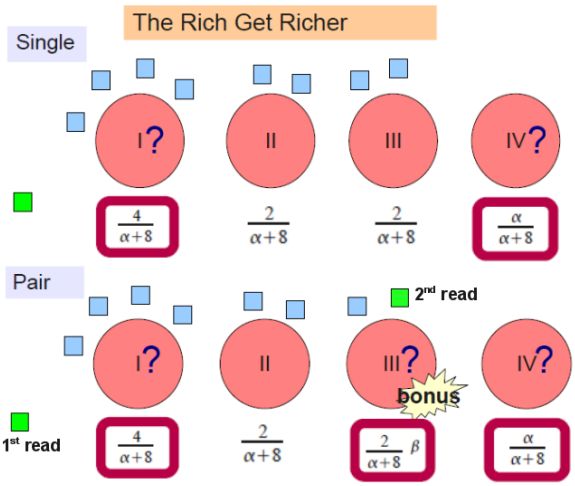

Fig. 3 Ilustration of CRP for Single and Paired end Read

pairs of reads in the contigs higher the likelihood, therefore using that procedure will decrease the likelihood of our model, opposite with the original Genovo which will increase its likelihood. Although not using that procedure, our model doesnt produce any chimera contig.

3.1 Modified Chinese Restaurant Process (CRP)

Genovo uses chinese restaurant processes to get prior accounts of the unknown number of genomes in the sample. The concept of CRP is the rich get richer, next customer sits at a table with probability proportional to number of customers already sitting at it and sits at new table with probability proportional to a concentration parameter, α . Most popular tables attract the most new customers and become even more popular [12]. In the assembly case, a customer is a read while a table is a contig. A CRP is a conditional distributions which is invariant to the order of the items which in our case are the reads [13]. The conditional distribution of CRP is represented by formula 9.

$$p(s_i = s|\mathbf{s}_{-i}) = \frac{1}{N-1+\alpha} \cdot \begin{cases} N_{-i,s} & s \text{ is an existing contig} \\ \alpha & s \text{ represents new contig} \end{cases} \quad (9)$$

$N_{i,s}$ counts the number of items, not including i, that are in contig s. For paired end reads, besides concerning with the concept of the rich get richer, it should also concerns that a pair of reads should be in a contig. Therefore, we give bonus if a read is in the same contig with its mate. In the illustration shown in **Fig. 3**, a read choose a contig. There are 3 contigs having reads and a new contig is possible to be created. In single read case, the contig which will be chosen depends on the number of read in the contig and the concentration parameter, α so that the candidate contigs are contig I (having the most read) and contig IV.

$$p(s_i = s|\mathbf{s}_{-i}) = \frac{1}{N-1+\alpha} \cdot \begin{cases} N_{-i,s} * \beta & s \text{ is a mate contig} \\ N_{-i,s} & s \text{ is an existing contig} \\ \alpha & s \text{ represents new contig} \end{cases} \quad (10)$$

While in paired end read case, it should also depend on the bonus parameter. Therefore the candidate contigs are contig I , contig III (having its mate) and contig IV. The conditional distribution of CRP in paired end read case is shown in formula 10.

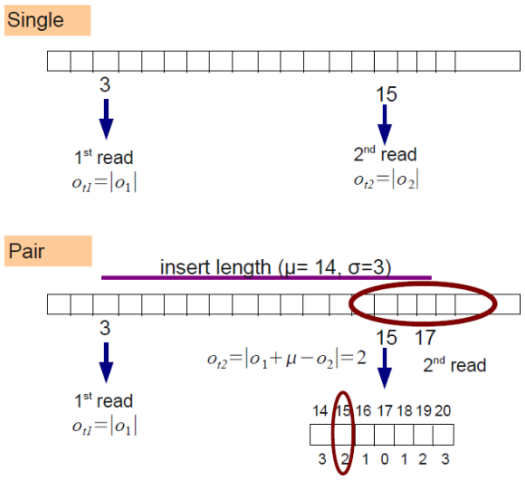

Fig. 4 Ilustration of CRP for Single and Paired end Read

Bonus parameter is represented by β. This bonus parameter intends a pair of reads should be in a contig so that can be one of the efforts to solve chimera case.

3.2 Modified Sampling Process

Sampling process means assigning a location of a read in the coordinate system of contigs and offsets. Geometric distribution represents the probability distribution of the number $y = x - 1$ of failures before the first success, shown in formula 11, p is the probability on each trial and k is the number of trial [14], [15].

$$P(x = k) = (1 - p)^k p \tag{11}$$

Genovo uses this concept. Sampling the beginning of a read (an offset) in a location x means that Genovo got failures sampling an offset in location 1 until $x1$ and success in location x. Genovo uses the negative and positive integer for the offsets representation in the contigs. A good contig is defined as a contig having the most reads toward the center of contigs. Therefore Genovo uses a symmetric variation of geometric distribution that includes all the negatives integers and is centered at 0 to sample a starting point oi of read within each contig, shown in formula 12.

$$G(o;\rho_s) = \begin{cases} 0.5(1-\rho_s)^{|o_t|}\rho_s & o \neq 0 \\ \rho_s & o = 0 \end{cases} \tag{12}$$

The number of trial, $|o_t|$, is the distance between offset and center (the absolute value of the offset). The parameter ρ_s controls the length of a contig. This parameter is same with the probability of success on each trial p in the original geometric distribution. As the posterior distribution of p can be determined if a Beta(α,β) prior is given [14], [15], Genovo also uses a known beta distribution to update the value of ρ_s. Genovo sets ρ_s to the mode of the Beta distribution $(1+N_s, 1+\beta+O_s) = \frac{N_s}{N_s+\beta+O_s}$ where $O_s = \sum_{K=1}^{N_s} |o_t|$

For paired end reads, the offset sampling process should care of the insert length parameter. In our modified, we use relative distance of the read to its mate considering the insert length. In the illustration shown in **Fig. 4**, there is a pair read with insert length distribution $(\mu, \delta) = (14, 3)$. Genovo uses the absolute value of offset as the number of trial, for an example in Fig. 4, the number of trial for the 1^{st} read is 3 and for the 2^{nd} read is 15. While in our modified, the number of trial for the 1^{st} read is same with Genovo, 3, yet we use relative distance for the 2^{nd} read. The relative distance is defined by $|o_1 + \mu - o_2|$. For an example in the Fig. 4, the number of trial $o_{t2} = |3 + 14 - 15| = 2$. The formula of symmetric geometric distribution for the 1^{st} read is same with Genovo shown in formula 13, while the ditribution for the 2^{nd} read is shown in formula 14.

$$G(o_1|\rho_{1s}) = \begin{cases} 0.5(1-\rho_{1s})^{|o_1|}\rho_{1s} & o_1 \neq 0 \\ \rho_{1s} & o_1 = 0 \end{cases} \tag{13}$$

$$G(o_{2s}|o_1, o_2, \rho_{2s}) = \begin{cases} 0.5(1-\rho_{2s})^{|o_{t2}|}\rho_{2s} & o_{t2} \neq 0 \\ \rho_{2s} & o_{t2} = 0 \end{cases} \tag{14}$$

where $o_{t2} = |o_1 + \mu - o_2|$

There is a possibility that a paired end read not sampled in the same contig. For this case, both the 1^{st} read and the 2^{nd} read are considered as 1^{st} read (single read). There are two ρ_s, ρ_{1s} for the 1^{st} read and ρ_{2s} for the 2^{nd} read. Both are updated using known Beta distributions. The ρ_{1s} is updated by the mode of distribution Beta$(1 + N_{1s}, 1 + \beta + O_{1s}) = \frac{N_{1s}}{N_{1s}+\beta+O_{1s}}$ where $O_{1s} = \sum_{k=1}^{N_{1s}} |o_1|$. The ρ_{2s} is updated by the mode of distribution Beta$(1+N_{2s}, 1+\beta+O_{t2s})$ $= \frac{N_{2s}}{N_{2s}+\beta+O_{t2s}}$ where $O_{t2s} = \sum_{k=1}^{N_{2s}} |o_{t2}|$. N_{1s} is the number of the 1^{st} read or single read (read which is not in the same contig with its mate) in contig s, o_1 is the offset of a read. N_{2s} is the number of the 2^{nd} read in contig s and o_{t2} is the number of trial for 2^{nd} read. By using this relative distance, a pair read sampled in the appropriate location in a contig has higher probability so that a contig produced is correct compared using default distance in Genovo.

3.3 Likelihood

The probability distribution in CRP and sampling process are changed so that the likelihood of the model changes. Same with Genovo, the likelihood of our model also consits of 4 components, shown in formula 4. The likelihood of the alignments $logp$(**x,y|s,o,b**) and the likelihood for generating (uniformly) each contig letter $logp$(**b**) are same with Genovo. While the differences are for the likelihood of contigs, shown in formula 15 and the likelihood of offsets, shown in formula 16, 17 and 18.

$$logp(\mathbf{s}) = S\,log(\alpha) + \sum_{i=1}^{S} log\Gamma(N_s) + log\Gamma(\alpha) - log\Gamma(N+\alpha) + N_{2s}log(\beta) \tag{15}$$

$$logp(\mathbf{o}|\mathbf{s}, \rho_{1s}, \rho_{2s}) = logp(\mathbf{o}_1|\mathbf{s}, \rho_{1s}) + logp(\mathbf{o}_2|\mathbf{s}, \rho_{2s}) \tag{16}$$

$$logp(\mathbf{o}_1|\mathbf{s}, \rho_{1s}) = \sum_{i=1}^{S} [O_{1s}log(1-\rho_{1s}) + N_{1s}log\rho_{1s} + N_{1s}log0.5] \tag{17}$$

$$logp(\mathbf{o}_2|\mathbf{s}, \mathbf{o}_1, \rho_{2s}) = \sum_{i=1}^{S} [O_{t2s}log(1-\rho_{2s}) + N_{2s}log\rho_{2s} + N_{2s}log0.5] \tag{18}$$

where S is the number of contigs, N_s is the number of read in contig s, N_{1s} is the number of 1^{st} read or single read in contig s, N_{2s} is the number of 2^{nd} read in contig s, $O_{1s} = \sum_{k=1}^{N_{1s}} |o_1|$ and $O_{t2s} = \sum_{k=1}^{N_{2s}} |o_{t2}|$. There is an additional for the likelihood of contigs which concerns about bonus parameter and 2^{nd} reads. In the likelihood of offset, there is two likelihoods, for the 1^{st} read and the 2^{nd} read. Same with Genovo, our modified also outputs the assembly that achieved the highest likelihood thus far.

Table 1 Simulated metagenomic dataset of 13 virus strain

	virus	LC(x)	MC(x)	HC(x)
1	Acidianus filamentous virus 1	10	10	5
2	Akabane virus segment L	5	5	5
3	Akabane virus segment M	5	5	5
4	Black queen cell virus	5	5	5
5	Cactus virus X	5	10	5
6	Chinese wheat mosaic virus RNA1	5	5	5
7	Chinese wheat mosaic virus RNA2	5	5	5
8	Cucurbit aphid-borne yellows virus	5	5	5
9	Equine arteritis virus	5	15	5
10	Goose paramyxovirus SF02	5	10	5
11	Human papillomavirus - 1	5	5	5
12	Okra mosaic virus	5	5	5
13	Pariacato virus chromosome RNA1	5	5	5

4. Results

We demonstrate the performance of our modified Genovo by comparing it with the original Genovo itself. We use MetaSim [16] to produce simulated metagenomic data of 13 virus strain which is used in Genovo‘s paper. We generate 50000 pair reads with length 250 bp and use the default 454 sequencing noise. The insert length distribution (μ, δ) is (3000,200).

We construct the dataset of different community complexity following [8], [17] (low, medium and high complexity, as LC, MC and HC respectively). The LC dataset has only one dominant organism, the MC dataset has some dominant organisms and the HC dataset has no dominant organism. In the HC dataset, we set 5 for the coverage of all virus. In the MC dataset, we set 15 for the coverage of one virus, 10 for the coverage of 4 virus and 5 for the others. In the LC dataset, we set 10 for the coverage of one virus and 5 for the others. List of the species and the coverage used in the dataset shown in **Table 1**.

Genovo set the parameter alpha α 2^{35}. This is the best parameter value to assemble for Genovo. To know the performance of our modified, we use some combinations of parameter between alpha α and bonus β. The combinations are α $\{2^{10}, 2^{20}, 2^{35}\}$ and β $\{1, 0.1\alpha, 0.3\alpha, 0.5\alpha\}$. The bonus $\beta = 1$ means that there is no bonus, because doesn‘t change the probability distribution in CRP. We run both our modified and Genovo for iteration from 25 until 200. As done in previous studies [4], [6], [18], we evaluate only contigs longer than 500 bp. Three measurements (total contig length, N50 and maximum length of contig) are used to evaluate the assembly capacity and two measurements (chimera rate and cover rate) are used to evaluate the assembly quality.

In the HC dataset, for each combination of parameter, our modified produces higher total length of contig than Genovo. If not using bonus parameter, our modified produces shorter N50 and max length of contig. It happens because the modified sampling process will be executed if a pair of read is in the same contig. Without bonus parameter means that there will be less pairs of reads in the contigs. Consequently, the modified sampling process less executed and the assembly process doesn’t work optimally. In each value, the best performance is for $\beta = 0.3\alpha$. The best performance of our modified is for a combination of $\alpha = 2^{20}$ and $\beta = 0.3\alpha$. The comparison between the best performance of our modified with Genovo shown in the left side of **Fig. 5**.

In the MC dataset, same with the HC dataset, our modified produces higher total length of contig than Genovo. If not using bonus parameter, our modified is as good as Genovo for $\alpha = 2^{10}$ but produces lower N50 with the increase of α. In each α value, the best performance is for $\beta = 0.3\alpha$. Same with the HC dataset, the best performance of our modified is for a combination of $\alpha = 2^{20}$ and $\beta = 0.3\alpha$. The comparison between the best performance of our modified with Genovo shown in right side of Fig. 5. In the first 25 iterations, our modified produces lower N50 and max length of contig, but it increases significantly with more iterations. In the iteration of 200 which reaches convergence, our modified is better than Genovo. The max length doesnt increase significantly.

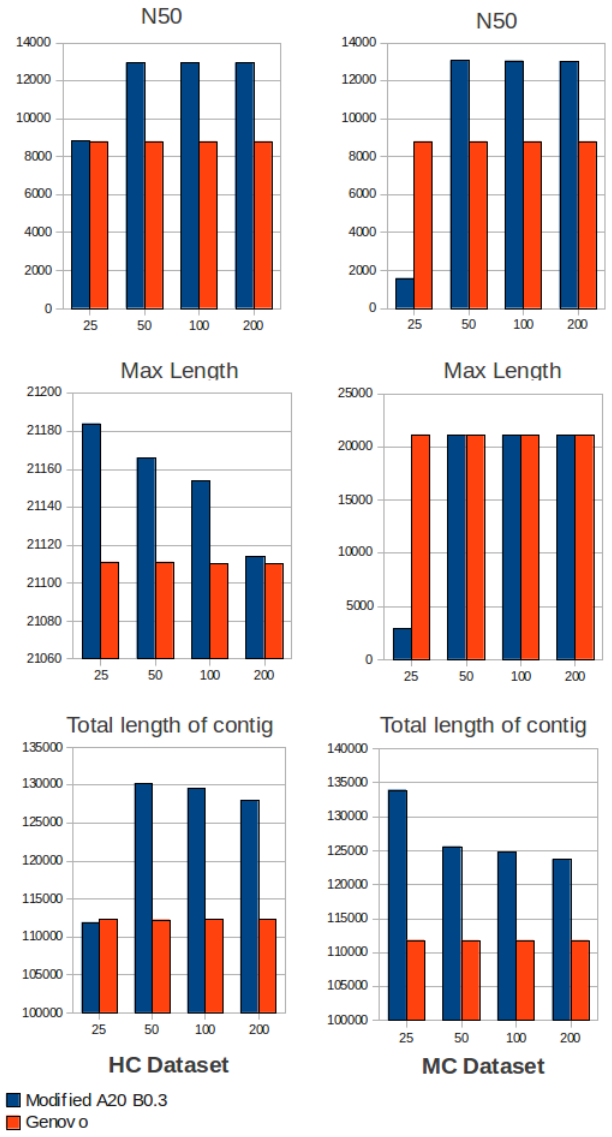

Fig. 5 The comparison of performances for HC and MC dataset

In the LC dataset, same with the other datasets, our modified produces higher total length of contig than Genovo. If not using bonus parameter, our modified is still as good as Genovo. In each value, the best performance is for bonus $\beta = 0.5\alpha$. The best performance of our modified is for a combination of $\alpha = 2^{35}$ and $\beta = 0.5\alpha$. The comparison between the best performance of our modified with Genovo shown in **Fig. 6**.

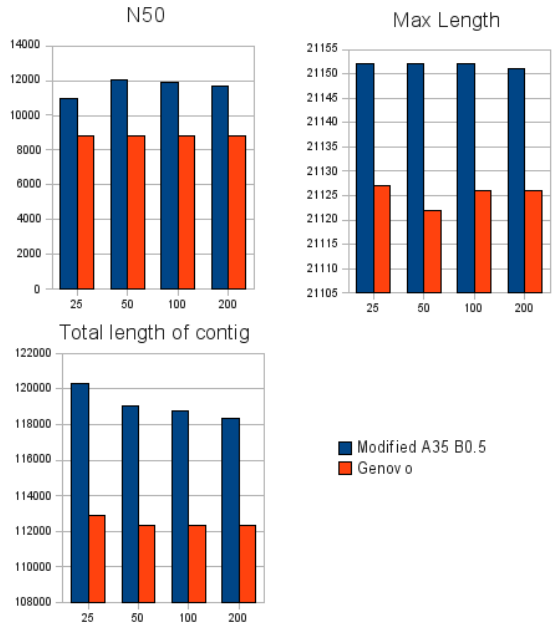

Fig. 6 The comparison of N50 and max length for LC dataset

The computational cost required for our modified is same with Genovo. For each dataset, both our modified and Genovo dont produce chimera contig. The cover rate of modified is better than Genovo but not significant, about 0.03.

5. Discussion and Conclusion

The main procedure of our modified model is the sampling process determining the location of a read in the coordinate system of contig and offset (the beginning of read). That is our reason why we attempt to modify CRP and offset sampling.

A CRP is a method to sample the contig. We add a bonus parameter in CRP which intends a pair of reads should be in a contig. When reaching convergence (in iteration 200), by using bonus parameter our modified produces assembly as good as, even better than Genovo for each dataset and each combination of parameter. For HC dataset, we can say that the performance decreases if not using bonus parameter. It means that bonus parameter which we propose gives good impact to the assembly process.

We use relative distance for the number of trial in symmetric geometric distribution in offset sampling process. Using this relative distance, a pair read which is sampled in the appropriate location in a contig based on the insert length parameter has higer probability. Therefore a read will be mapped in correct location. In each dataset for every combination of parameter, even not using bonus parameter, our modified produces higher total length of contig or can be said assembles more reads. This result means that this relative distance strategy can work well. Our modified doesnt increase the computational cost, same with Genovo. Based on the result, it can be concluded that our strategies can work well achieving better performance for paired end read.

Although our strategies give better results for paired end read, there are several works left in order to develop a great metagenomic assembler. We are going to continue our research. The next step, we are going to develop a proper scaffolding procedure. There are several challenges to scaffold for metagenome case. The main challenge is to develop an assembler with scaffold procedure which can automatically generate contiguous assemblies yet accurately capture genomic variation information throughout the assembly process [19]. Mate pair information is useful to solve this case [2], [7], [8], [18], [19], [20]. Short read, for example Illumina reads, have been gaining popularity, even for metagenomic studies [4]. We are going to extend this model for short read data. Actually, this model can be implemented for short read data, but the computational cost required is high enough compared to the other metagenome assemblers. We are going to solve this case in order to develop a great metagenome assembler with a realiable compuational cost.

References

[1] Scholz M.B., Chi C., Chain P.SG.: Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis, *Curr Opin Biotechnology*, Vol. 11.013 (2011)
[2] Peng, Y., Leung, H.C.M. , Yiu, S.M., Chin, F.Y.L.: Meta-idba: a de novo assembler for metagenomic data, *Bioinformatics*, Vol. 27(13), pp.i94-i101 (2011)
[3] Hiatt JB, Patwardhan RP, Turner EH, Lee C, Shendure J.: Parallel, tag-directed assembly of locally derived short sequence reads, *Nat Methods*, Vol. 7(2), pp.119-122 (2010)
[4] Qin, J., Li, R., Raes, J., et al.: A human gut microbial gene catalogue established by metagenomic sequencing, *Nature*, Vol. 464, pp. 59-65 (2010)
[5] Namiki T., Hachiya T., Tanaka H., and Sakakibara Y.: MetaVelvet : An extension of Velvet assembler to de novo metagenome assembly from short sequence reads, *Nucleic Acids Res.*, in press (2011)
[6] Laserson J., Jojic V. and Koller D.: Genovo: de novo assembly for metagenomes, *J Comput. Biol.*, Vol. 18, pp.429-443 (2011)
[7] Peng, Y., Leung, H.C.M. , Yiu, S.M., Chin, F.Y.L.: IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth, *Bioinformatics*, Vol. 28 (11),pp.1420-1428 (2012)
[8] Lai B., Ding R., Li Y., Duan L., Zhu H.: A de novo metagenomic assembly program for shotgun DNA reads, *Bioinformatics*, Vol. 28(11), pp.1455-1462 (2012)
[9] Wommack,K.E. et al.: Metagenomics: read length matters, *Appl. Environ. Microb.*, Vol. 74, pp.14531463 (2008)
[10] Lasken, R.S., Stockwell, T.B.: Mecanism of chimera formation during the multiple displacement amplification reaction, *BMC Biotechnology*, Vol. 7:19 (online)
DOI: :10.1186/1472-6750-7-19 (2007)
[11] Ewing, B., Green, P.: Analysis of Expressed Sequence Tags Indicates 35,000 Human Genes, *Nature Genetics*, Vol. 25, pp.232-234 (2000)
[12] Johnson M.:Chinese Restaurants Process(online),
available from ⟨http://cog.brown.edu/ mj/classes/cg168/slides/ChineseRestaurants.pdf⟩ (accessed 2012-06-01).
[13] Aldous, D.J.: *Exchangeability and related topics*, Lecture Notes in Math, Vol. 1117, pp. 1198 Springer, Berlin (1985)
[14] Casella G., Berger R.L.: *Statistical Inference*, Duxbury Press, 2nd edition (2001)
[15] Hoog, R.V., Tanis, E.: *Probability and Statistical Inference*, Pearson Publisher, 8/E (2010)
[16] Richter D.C., Ott F., Auch A.F., Schmid R., Huson D.H.: MetaSimA Sequencing Simulator for Genomics and Metagenomics, *PLoS ONE*, Vol. 3 (10), e19984 (online)
DOI: :10.1371/journal.pone.0003373 (2008)
[17] Pigmatelli M., Moya A.: Evaluating the fidelity of de novo short read metagenomic assembly using simulated data, *PLoS ONE*, Vol. 6, e19984 (online)
DOI: :10.1371/journal.pone.0019984 (2011)
[18] Chaisson, M.J., and Pevzner, P.A.: Short read fragment assembly of bacterial genomes, *Genome Res.*, Vol. 18, pp. 324–330 (2008)
[19] Koren S., Treangen T.J. and Pop M.: Bambus 2: Scaffolding Metagenomes, *Bioinformatics*, Vol. 27(21), pp.2964-2971 (2011)
[20] Chaisson, M.J.P., Brinza, D., and Pevzner, P.A.: De novo fragment assembly with short mate-paired reads: does the read length matter? *Genome Res.*, Vol. 19, pp. 336-346 (2009)