

# A Gibbs Sampling Approach to Diploid Genome Reconstruction

Jong Hyun Kim,<sup>1 2</sup> Lei Li,<sup>1</sup> Michael Mehan,<sup>1</sup> Michael S. Waterman<sup>1 3</sup>

**Keywords:** computational genomics, polymorphism, haplotype, the Gibbs sampler

## 1 Introduction

Genome sequencing assemblers have produced the haploid genome sequence irrespective of the fact that true target is diploid in an eukaryotic genome. Therefore, it is often a challenge to assemble the genomes of eukaryotic organisms, in particular of highly polymorphic organisms. We previously proposed a statistical reconstruction method to address this problem[1]. In this poster, we present a novel approach, the Gibbs sampler, for identifying polymorphisms and inferring haplotypes from the identified polymorphisms. In this approach, the target genome is diploid. The mate-pair information from end-sequenced clone inserts is exploited to provide long-range linkage. This algorithm also estimates the missing origins of shotgun reads. The Gibbs sampler is applied to the genome assemblies of two highly polymorphic organisms, *Ciona intestinalis* and *fugu rubripes*. The draft genomes and all the shotgun reads of two organisms are publicly available at [genome.jgi-psf.org/ciona4](http://genome.jgi-psf.org/ciona4) and [genome.jgi-psf.org/fugu](http://genome.jgi-psf.org/fugu). Currently, their polymorphism rates are reported to be 1.2% and 0.4%, respectively [2][3]. The polymorphism rates based on our approach are reported. The accuracy of the estimation is evaluated on the simulated dataset.

## 2 Method

In sequence assembly, layout in the region of interest can be thought to be a  $m \times n$  matrix. In this matrix, we denote base-calls by  $X = \{X_{ij}; i = 1, 2, \dots, m, j = 1, 2, \dots, n\}$ . Similarly, we denote the haplotypes by  $S = \{S_{kj}; k = 1, 2, j = 1, 2, \dots, n\}$ . Let  $F = \{F_i; i = 1, 2, \dots, m\}$  be the origins of reads; That is,  $F_i$  is 1 if the read comes from haplotype 1, and  $F_i$  is 2 if the read comes from haplotype 2. In MCMC approach, a time variable  $t$  is introduced to these random variables. Therefore,  $S_{ij}^{(t)}$  is  $S_{ij}$  at time  $t$ , and  $F_i^{(t)}$  is  $F_i$  at time  $t$ . Let  $S_{[-A]} = \{S_{ij}; (i, j) \in A^c\}$  and  $F_{[-A]} = \{F_i; i \in A^c\}$ .

The following two steps are repeated until  $t$  reaches the maximum  $T$ . Samples are discarded as *burn-in* until  $(s^{(t)}, f^{(t)}) = (s^{(t-1)}, f^{(t-1)})$  condition is satisfied. The chain is *thinned* by storing the sample from every  $k$ -th iteration.

1. For each  $\{(i, j); i = 1, 2, j = 1, 2, \dots, n\}$ , draw  $s_{ij}^{(t+1)}$  from  $P(S_{ij}^{(t+1)} | S_{[-(i,j)]}^{(t)}, F^{(t)} = f^{(t)}, X)$  and set the remaining components as  $s_{[-(i,j)]}^{(t+1)} = s_{[-(i,j)]}^{(t)}$ .

---

<sup>1</sup>Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA.  
E-mail: {jonghkim, lilei, rielmeha, msw}@usc.edu

<sup>2</sup>Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA.

<sup>3</sup>Informatics Research, Celera Genomics, 45 West Gude Drive, Rockville MD 20850, USA.

2. For each  $\{i; i = 1, 2, \dots, m\}$ , draw  $f_i^{(t+1)}$  from  $P(F_i^{(t+1)} | F_{[-i]}^{(t)}, S^{(t+1)} = s^{(t+1)}, X)$  and set the remaining components as  $f_{[-i]}^{(t+1)} = f_{[-i]}^{(t)}$ .

The details of reconstruction algorithm are omitted here.

### 3 Dataset

Low-quality regions of shotgun reads were trimmed by using LUCY[4]. The quality-trimmed shotgun reads were then aligned to the published *ciona intestinalis* draft genome by using BLASTN[5]. In each pairwise alignment, the distance between two end-sequenced reads was also considered. Only the reads uniquely mapped to the reference genome were included in the layout. The alignments spanned  $\sim 95\%$  of the draft genome of which size is 116.7 Mbp. The shotgun coverage was  $\sim 8$ , which is quite consistent with the reported coverage[2].

### 4 Preliminary Results

Polymorphisms are identified in a scaffold (scaffold 990, size: 16 Kbp), and its polymorphism rate was  $\sim 1.6\%$ , which is higher than the overall rate reported[2]. The accuracy of the estimation was validated through simulation. Using the scaffold layout and the draft genome sequence, we generated polymorphic sites according to the reported rate[2]. In this simulation, sequencing error rate was based on the real quality score of each base, and true positive rate was  $\sim 99\%$ . Full results and analyses will be presented elsewhere.

### 5 Acknowledgements

This work is supported by the NIH CEGS grant.

### References

- [1] Li, L., Kim, J. H. and Waterman, M. S. 2004. Haplotype Reconstruction From SNP Alignment. *Journal of Computational Biology*, 11, pp. 505-516.
- [2] Dehal, P. et al. 2002. The draft genome of *ciona intestinalis*: insights into chordate and vertebrate origins. *Science*, 298, pp. 2157-2167.
- [3] Aparicio, S. et al. 2002. Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu rubripes*. *Science*. 297, pp. 1301-1310.
- [4] Hui-Hsien Chou and Michael H. Holmes 2001. DNA sequence quality trimming and vector removal *Bioinformatics*. 17, pp. 1093-1104.
- [5] Altschul, S. F. et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search program. *Nucleic Acids Research*. 25, pp. 3389-3402.