

Second Generation DNA Alignment Tools

Christie Robertson(1), Eric Flynn(1), Gary Montry(2), Sandra Porter(1), Todd Smith(1)

1) Geospiza, Inc. Seattle, WA www.geospiza.com 2) Southwest Parallel Software, Albuquerque, NM www.spssoft.com



Southwest Parallel Software
Providing High-Performance Software Solutions



1 Abstract

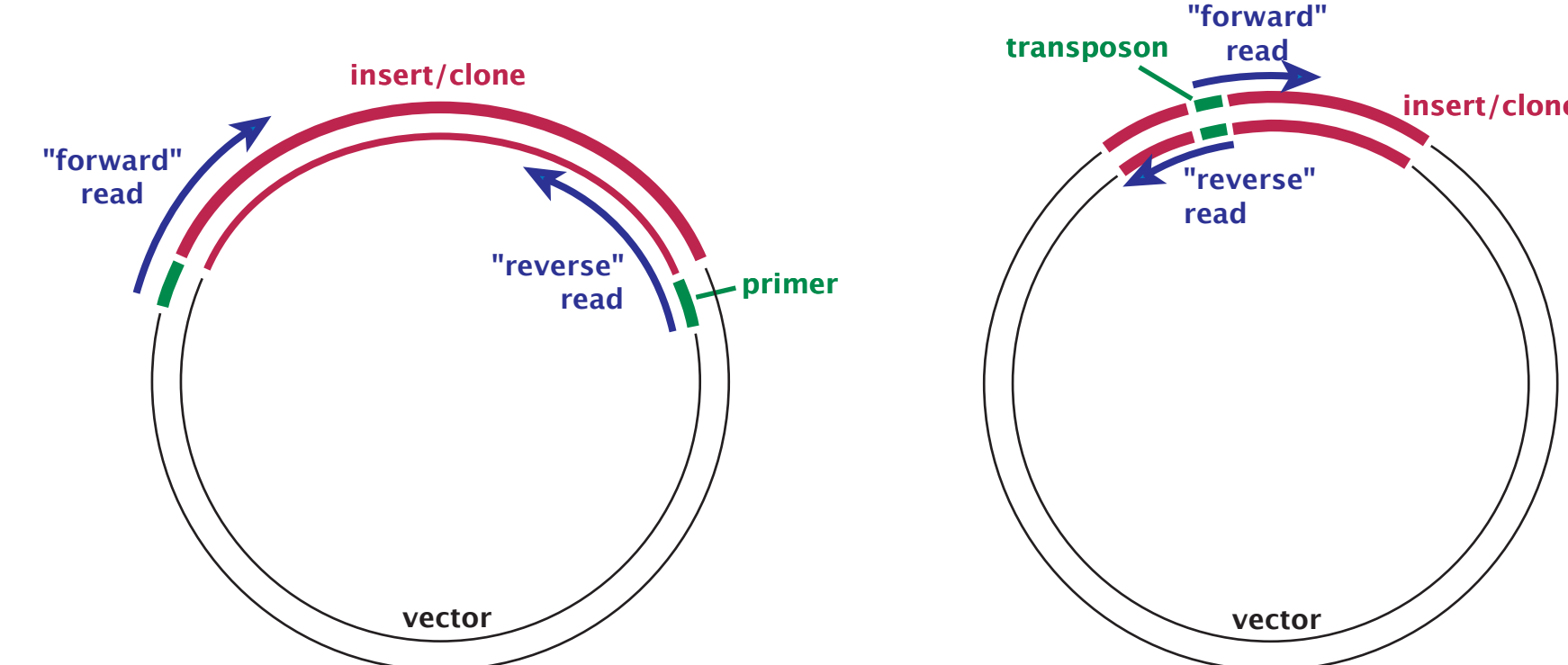
The human genome project spurred the development of high throughput technologies, especially in the area of DNA sequencing. Not only has this effort uncovered the sequence of the human genome, it has catalyzed development of an entire industry based on DNA sequencing and genomics. Since these technologies produce enormous amounts of data, they depend on bioinformatics programs for data management. Phrap, Cross_Match, RepeatMasker, and Consed have played an integral role in genome projects and have come to be accepted as standard tools for genomic alignment and assembly. As sequencing technology and software have evolved, however, so too have the scientific applications that rely on these programs. Specific needs associated with whole genome shotgun sequencing, EST cluster analysis, and genotyping applications highlight the importance of updating standard bioinformatics programs to meet the requirements of a broader community. We are re-engineering Phrap, Cross_Match and RepeatMasker to improve their performance and utility through optimizing the core algorithms and developing a framework to store, manipulate, and view assembled sequence data. We are developing a structure through which specific XML-formatted hints and constraints will be able to pass instructions to the core alignment program, giving it information on the handling of parts of the data, or the data set as a whole, in individualized ways. Hints regarding read pairs, associations or non-associations between reads or contigs, sequencing reaction conditions, highly-repetitive regions, reference sequences, and other information will be able to be applied to direct sequence alignment, without altering the underlying data itself. In addition, a new viewing program is being developed to review, edit, and manipulate sequences, giving users unprecedented control over their data.

2 DNA assembly problems

Assembly programs use algorithms to align nucleotide sequences to each other based on similarity between the sequences. Every algorithm has to make choices as to thresholds at which sequences are similar enough to be aligned to each other and when they are dissimilar enough to be separated. It is an inevitability that any algorithm will wrongly fail to assemble in some cases and wrongly assemble in other cases, and every algorithm has to choose its demons in this regard. The popular DNA assembly program Phrap (www.phrap.org, University of Washington-Seattle), which utilizes a dynamic programming algorithm based on Smith-Waterman (Smith and Waterman, 1981), exhibits fairly good accuracy in aligning sequences into appropriate groupings, but like all assembly algorithms, there is also a good deal of misassembly (Semple et al., 2002). Highly-repetitive regions of DNA, such as those that are rampant in the human genome (while the Human Genome Project is considered "completed finished" a large percentage of it is unlikely to be accurately deciphered in the near future, due to its highly-repetitive nature (Lander et al., 2001)), are one source of assembly problems. Another challenge is caused by the application of assembly algorithms to increasingly wide arrays of scientific problems, such as EST clustering, genotyping, and comparative genomics.

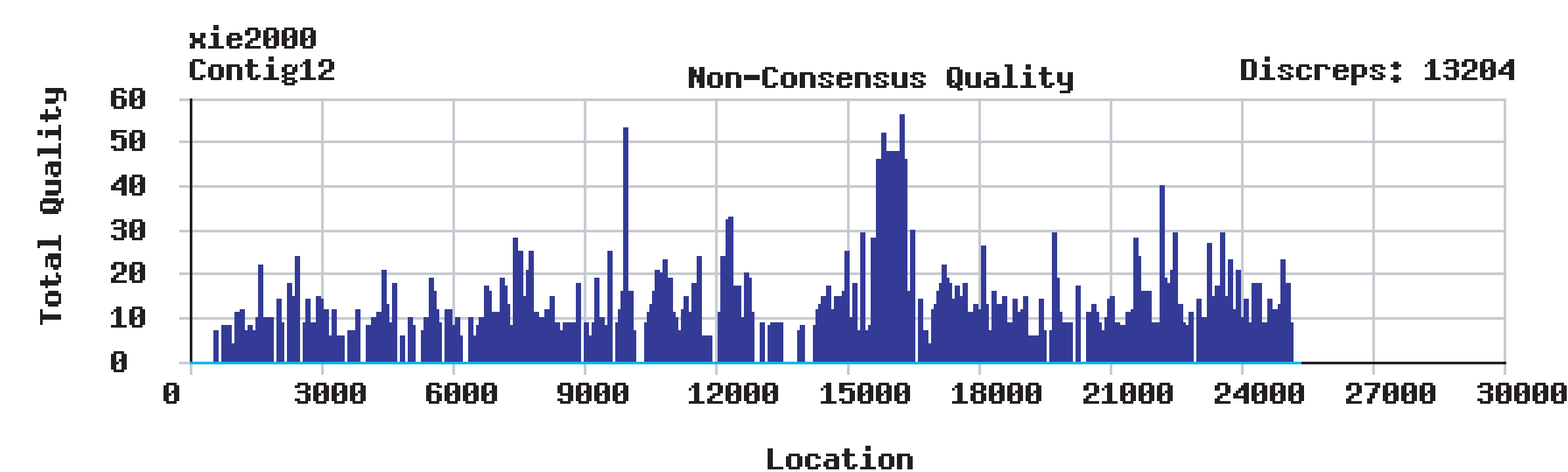
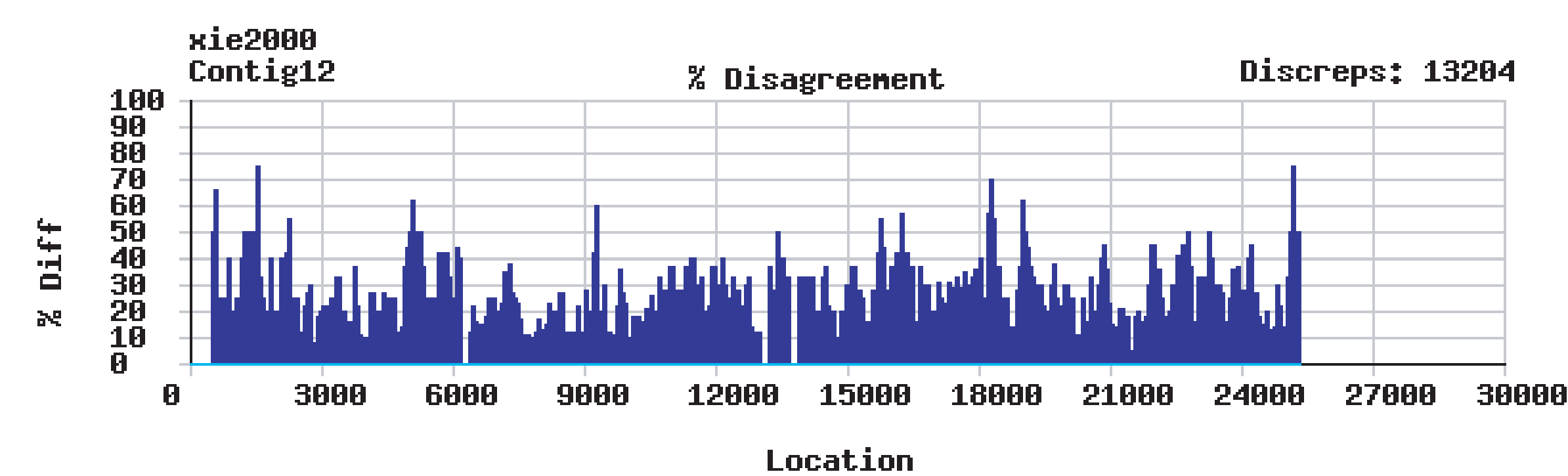
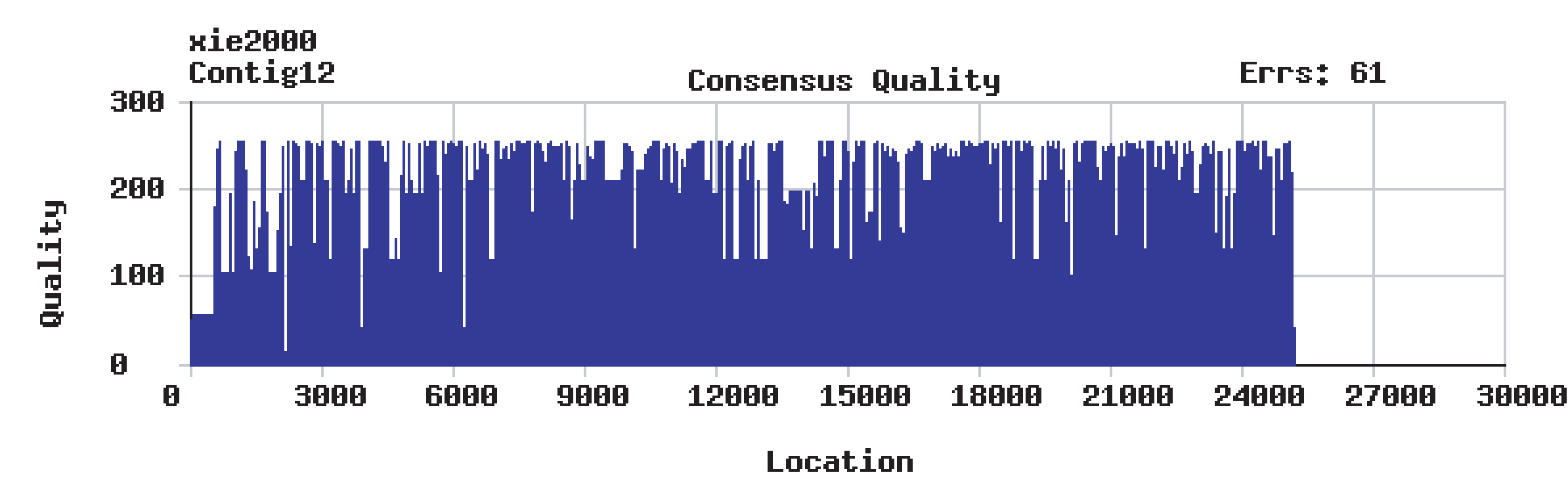
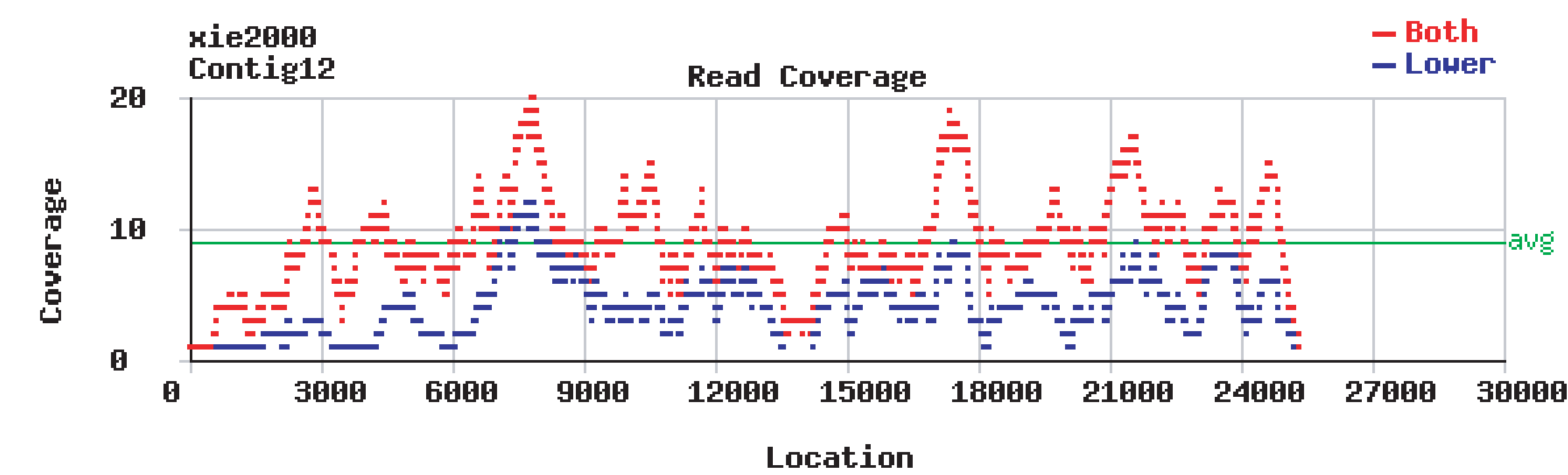
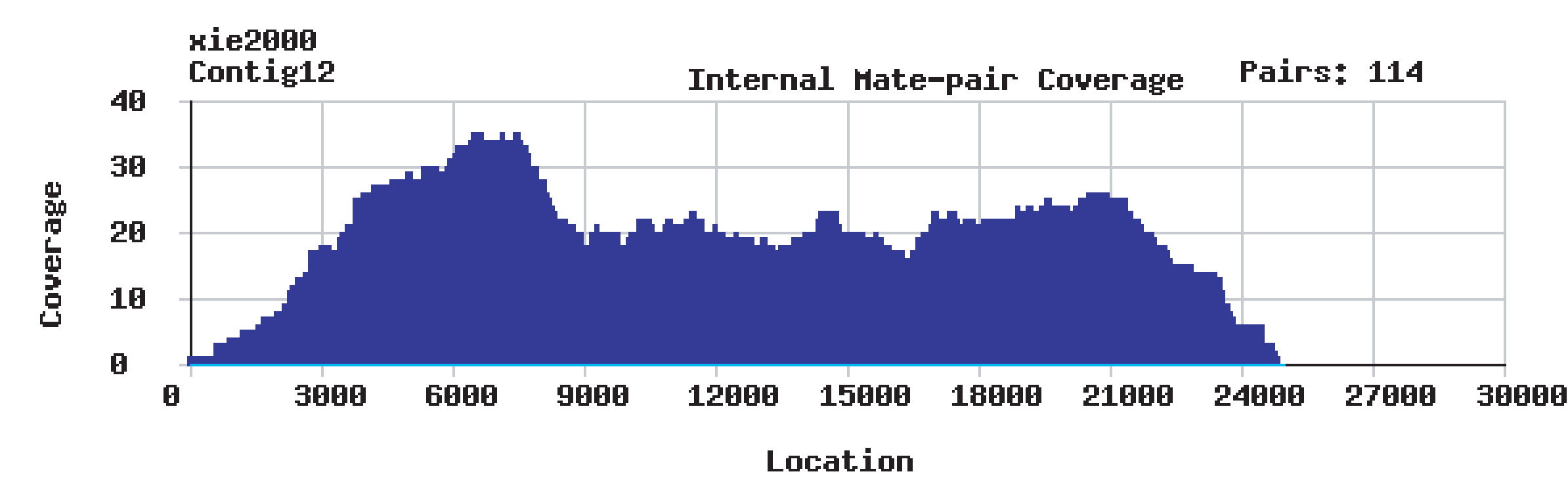
3 Assisting the assembly algorithm

An elegant algorithm is a good core for alignment, but because of the variety of problems and applications posed by real biology, the performance of a core assembly algorithm can be greatly enhanced by specific information that a scientist knows about his or her data set. We are building an assembly program based on Phrap, but with the ability to apply additional "hints" to the assembly process. Mate pairs (AKA read pairs, forward/reverse pairs) will be of key assistance to the assembly process.



4 Feedback regarding misassembly

Hints regarding expected read coverage, mate pairs, and expected stringency can be used to identify regions where alignment has occurred in error or where an appropriate alignment has likely been missed. The graphs below represent statistics from mouse genomic sequencing data assembled with an early version of our re-engineered algorithm. Finished sequences from BAC M_BB0159J17 were kindly provided by Pat Minx and Rick Wilson at Washington University, St. Louis.



5 Development of an alignment program

Beginning with a core alignment algorithm based on Phrap, we are developing an alignment program that will allow application of "hints" to the assembly process and the analysis of assembly success.

Key aspects of the project include:

- transfer and storage of input and output information in the NCSA's HDF (high density format), allowing swift, accurate access to the information therein
- a graphical user interface for controlling and viewing assemblies
- "project modes" that allow an assembly to be tuned for a specific situation (EST clustering, genotyping, shotgun sequencing, repetitive DNA)
- the ability to apply information about a subset of the assembly (mate pairs and other read-to-read anchoring hints, scaffold mapping information, desired stringency of alignment, stringency of alignment, highly-repetitive character) to the assembly process

6 Current Re-engineering Progress

- Redesigned Phrap's contig handling and added a new set of superstructures to build ordered and oriented maps.
- Redesigned Phrap's repeat handling routine to better use minmatch, maxmatch and minscores parameters.
- Added an identity variable to detect repeats during assembly and improve the ability to reject repeats during merge processes.
- Completed initial development to use mate pairs to guide assembly.

7 References

- Gordon, D., Desmarais, C., and Green, P. (2001). Automated finishing with autofinish. *Genome Res* 11, 614-625.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.
- Semple, C. A., Morris, S. W., Porteous, D. J., and Evans, K. L. (2002). Computational comparison of human genomic sequence assemblies for a region of chromosome 4. *Genome Res* 12, 424-429.
- Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol* 147, 195-197.

8 Acknowledgements

We gratefully thank Pat Minx (Washington University, St. Louis) for his expert guidance regarding misassembly problems caused by repetitive DNA sequences. We also thank Pat Minx and Rick Wilson (Washington University, St. Louis) for the contribution of finished mouse BAC sequence. We thank the National Institutes of Health for SBIR grant #R44 HG02244-02, which supports this project.

Geospiza, Inc. and the finch logo are registered trademarks of Geospiza incorporated. All other trademarks, service marks and registered trademarks appearing herein are the properties of their respective owners and are hereby acknowledged. Presented at Recomb-Satellite, Palo Alto CA, 2003