PublisherInfo		
PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

Two's twice as useful

ArticleInfo		
ArticleID	:	3918
ArticleDOI	:	10.1186/gb-2001-2-5-reports0013
ArticleCitationID	:	reports0013
ArticleSequenceNumber	:	15
ArticleCategory	:	Paper report
ArticleFirstPage	:	1
ArticleLastPage	:	4
ArticleHistory	:	RegistrationDate : 2001–4–9 Received : 2001–4–9 OnlineDate : 2001–5–10
ArticleCopyright	:	BioMed Central Ltd2001
ArticleGrants	:	

ArticleContext	:	130592255

Jonathan B Weitzman

Abstract

Comparison of the two drafts of the human genome sequence has revealed reassuring similarity but also some differences.

Significance and context

The race to sequence the human genome has resulted in two independent drafts of the genome sequence, generated in different ways. The public Human Genome Project (HGP) sequence was assembled using a systematic, hierarchical mapping and sequencing approach, whereas the private Celera Genomics draft is the result of a whole-genome random-shotgun approach. Differences between the experimental approaches and the conditions for access to the genome data have been hotly debated. Now that the genome race has been officially declared a draw, the two sequences offer a rich resource for detailed computational analysis of the molecular context and architecture of the human genome. The availability of two sequences (each of which was generated from DNA representing pooled samples from several individuals) enables a comprehensive analysis of the human genome and highlights the challenges ahead.

Key results

Aach *et al*. used computational analysis to compare the quality and content of the two draft sequences. They compared 2.9 gigabases (Gb) of merged data from the public database (nonredundant, HGP-nr) and 2.9 Gb of the Celera database (Cel). At first glance the two drafts are reassuringly similar in length and content. Aach and colleagues report that differences can be revealed by detailed computational analysis. Many of these differences are likely to diminish as completion and annotation progress. Analysis of the quality of the drafts showed that HGP-nr contains fewer unidentified bases than Cel (0.65% versus 8.7%). The authors analyzed ten of the longest genes to evaluate the continuity of the assemblies. The two drafts had comparable limitations in this respect - Aach *et al*. found six genes in HGP-nr and seven in Cel with both ends on the same continuous contigs. They analyzed the frequency of all random 15-nucleotide sequences to determine candidate unique 15mer sequences (cu15s). For each draft genome they found over 160,000 cu15s of which about 11% are not shared between the two drafts (this figure is reduced to 0.14% after adjusting for the predicted rate of false negatives). Aach *et al*. used weight matrices to assess the frequency of sequence motifs that bind DNA-binding proteins

within sequences 4 kilobases (kb) upstream of 3,352 genes. They found that upstream sequences were significantly enriched for some binding-site motifs (that for the EGR-1 zinc finger protein, for example) but not for others (that for the CRX photoreceptor homeobox factor, for example).

Links

Further information about the human genome sequence can be found from the National Center for Biotechnology Information and from Celera Genomics. Additional data from this paper and bioinformatics resources are available from the Lipper Center for computational genetics maintained by the Church laboratory.

Conclusions

The authors highlight the limitations of the first drafts of the human genome. The two assemblies are similar in respect of size, unique sequences and frequency of binding motifs in the DNA. Improvements in annotation and completion will remove some of the current limitations. The sequence data provide a challenge to computational analysis as increased bioinformatics resources will be needed, and the data will be undergoing constant updating. The authors predict that these problems will be rapidly addressed by the development of algorithms and dedicated informatics platforms.

Reporter's comments

The more complete and high-quality genome sequences that are available, the more will comparative analysis of this type be able to provide insight into the differences between humans and other mammals, as well as the nature of genetic differences between individuals. There may prove to be additional benefits from having two human genome sequences rather than one. The greatest challenge will undoubtedly be to integrate insights from such computational analysis into biologically relevant models.

Table of links

Nature

National Center for Biotechnology Information

Celera Genomics

Lipper Center for computational genetics

References

1. Aach J, Bulyk ML, Church GM, Comander J, Derti A, Shendure J: Computational comparison of two draft sequences of the human genome. Nature. 2001, 409: 856-859. 0028-0836

This PDF file was created after publication.