Comparison of Draft Human Genomes

Ian O'Dwyer

November 16, 2001

Earlier this year, a major breakthrough in biology was announced: the sequencing of the human genome. Two separate groups, The Human Genome Project[1] (HGP) and Celera Genomics[2], independently produced a draft genome and work continues to complete the sequencing. Although the current genome is only considered *draft* at this stage, it has opened up an enormous range of exciting possibilities for new research ranging from cures for cancer to understanding the genetic causes of drug addiction. Against this background of excitement and enthusiasm for the sequencing of the human genome, it is interesting to step back and objectively compare the two versions of the genome, their level of completeness and the work which still needs to be done in order to bring the sequence to 'complete' status from the current draft. We begin with a discussion of why the current genome can only be considered a draft version and then look at the differences and similarities between the two genomes and the methods which generated them. Much of this analysis follows that in the paper by Aach et al[3].

Currently, no eukaryotic organism has been sequenced to 100%. Some repetitive regions of the genome are particularly difficult, perhaps impossible, to clone. Cloning is one of the first steps in the sequencing process and is discussed further below. Researchers anticipate that these regions will not contain a significant proportion of protein coding genes and so their omission may not result in a significant loss of information in the genome. However, it seems difficult to state with any degree of certainty that regions of the genome are unimportant if they cannot be decoded and analysed. After all, until recently it was thought that the human genome would contain as many as 100,000 genes. After sequencing it seems that this number will be closer to 35,000. Until the genome was sequenced there was not a strong concensus that the human genome would contain barely twice as many genes as that of the common earth worm, so to discount unsequenced regions as unimportant seems somewhat premature. On the other hand, there is some evidence that certain regions of genomes in other species do contain the majority of the genes. For example, *Drosophila Melanogaster*, the fruit fly, had only about two thirds of its genome sequenced, but 97% of the euchromatic portion was sequenced and this is believed to be where the majority of the genes reside[4]. Since each species varies widely in the amount of its genome which can be sequenced, a judgement has to be made as to when the sequence is considered complete. In the case of the human genome, the stated goals are to obtain a sequence in which less than one base in 10,000 is incorrectly placed, more than 95% of the euchromatic region is sequenced and each gap is smaller than 150 kilobases. Under these guidelines, only about one quarter of the HGP and Celera genomes can be considered complete, although the Celera version is a little more 'gappy' than the HGP version. The *draft* status comes from the fact that the sequences still have many gaps in them. With too many gaps it becomes hard to align and orient all the small strings of sequence which make up the genome and are the product of the early stages of the sequencing process. So in this sense, the two sequences must indeed be considered

draft genomes only, with a great deal of work required to finish off the sequence. This work is likely to proceed much more slowly than the initial sequencing since it is reasonable to assume that the easiest parts of the genome to sequence have already been completed, and those sections which remain are harder to sequence correctly using current techniques.

HGP and Celera used quite different strategies in order to sequence the human genome. Both approaches utilise what is known as *shotgun sequencing*. The basic premise is to spray the DNA strand to be sequenced with a series of small sequence reads, much as shotgun pellet sprays a target. Each of these small sequences then overlaps at some level and ultimately yields the complete sequence of the genome. However, the two groups chose to apply this technique in a very different manner. HGP initially generated a series of clones (intermediate-sized sequence fragments copied from the genome) which cover the entire genome and provide a degree of overlap and redundancy, as discussed above. Reconstruction of the genome sequence is then performed by considering the overlap, mapping and chromosomal information for each of the clones. In contrast to this, Celera used a shotgun approach for the entire genome, without first generating a series of clones. One major consequence of this is that the Celera genome contains larger gaps in terms of the number of consecutive unidentified bases than does the HGP genome, although there are some caveats to this discussed below.

Comparison of the draft genomes produced using the above methods was performed by Aach et al[3] using three versions of the genome. The first version, denoted HGP-all, contains 34,084 large insert clones taken directly from the HGP data and contains 4.8 Gigabases (Gb). This version is highly redundant as it contains many overlapping clones and other sources whose overlaps are not well defined. The second version, HGB-nr, the nr indicating no redundancy, contains 2.9Gb from 6,094 sequences and represents a refinement of HGP-all. All clearly identifiable redundancy has been removed, hence the much smaller number of bases and sequences. The final version of the genome, Cel, is taken from Celera's *Human Genome D*, representing 2.9Gb in 54061 sequences.

Overall, HGP-nr has fewer unidentified bases than Cel, with 0.65% vs 8.6%, however this is partly to do with the way gaps are represented differently between the two sequences. Gaps in the sequence are one of the primary sources of concern in determining the completeness of the genome. Both groups use N's to denote gaps, although they are used in a slightly different way by each. Cel uses an N for each unknown base in the sequence, regardless of the size of the gap. HGP-nr and HGP-all use the same policy up to gaps of 100 bases, but after this, the number of N's is generally (but not always) kept at 100 regardless of how large the gap is. Thus when considering figure 1., it is important to remember that the smaller continuous strings of N's in HGP does not imply that the gaps are smaller than Cel. Cel contains a total of 169,779 stretches of N's (i.e. gaps) ranging in size from 1 to 168,735 bases. HGP-nr contains 407,686 gaps ranging from 1 to 2,500 bases. The gaps in the sequence appear to make a statistically significant comparison of the two genomes extremely troublesome. Since the gaps occur at different locations and are different sizes in the two cases, any attempt to align the genomes appears to be an educated guess at best.

As a test of continuity of the genomes, the 10 genes with the largest m-RNA in the RefSeq database were selected and a test was performed to see whether they could be identified in the draft genome by the BLAST search algorithm. This is essentially a test of the longest contiguous sections of genome (contigs) that HGP and Celera could generate by their shotgun process compared with the longest contiguous sections which nature is capable of transcribing. The results showed, perhaps not unexpectedly given the current level of technology, that both groups have a limited ability to produce contigs in their genomic sequences as long as those produced in biological organisms.

Short strands of DNA consisting of two to twenty bases (oligonucleotides) which occur only once in the genome sequence were used to give a statistical view of how much sequence content the two genomes have in common. Aach et al used stretches of 15 nucleotides, 15-mers, and point out that there are about 4^{15} such sequences in a 6Gb genomic sequence (considering both strands). The number of occurences of such a 15-mer follows a roughly poisson distribution with mean $\sim$6. Every possible 15-mer was considered and a computer algorithm was used to determine if the 15-mer occured 0, 1, or multiple times in the genome. Clearly this process is somewhat flawed since there are many gaps in the sequence which could cause the 15-mer to be recognised multiple times when, in fact, it exists only once and one can imagine many other such scenarios which would cause equivalent problems. However, it seems to provide a sensible test of at least those portions of the two genomes which are not too 'gappy'. The net result of the analysis was that $\sim$160,000,000 15-mers were found in both sequences, with about 11% of them not being shared between the HGP and Celera sequences. This results in about 0.14% of the sequence being present in one database and not the other. It therefore appears that both genomes contain about the same amount of unique sequences and that the majority of unique sequences are shared by both databases.

In summary, the two draft genome sequences from Celera and HGP appear to be quite similar at a superficial level but differ in the details. Superficially, the sequences contain similar numbers of nucleotides, have comparable amounts of unique sequence and appear to share most of these sequences, as shown by the quasi-stastical approach outlined above. However, the methods used to form the sequence lead to some critical differences. The amount of contiguous sequence in the two cases is quite different both in terms of size and gap distribution. Whilst the total amount of unidentified bases are similar, the HGP genome contains a higher number of small gaps whereas the Celera sequence has more long gaps. Also, due to the HGPs use of small clones to build up the genome, they are able to present four stages of sequence data with a higher degree of annotation compared with the single Celera Human Fragments database. Presumably the Celera group will work to repair some of these slight deficiencies compared to the HGP and both sequences should become more complete and detailed. Without doubt, the century of biology has got off to a roaring start and the complete elucidation of the human genome is fast becoming science fact rather than science fiction. Much hard work remains and new techniques will need to be developed in order to decode the more tricky parts of the genome, but the potential rewards are vast, both financially and altruistically. This will ensure that there is no shortage of resources, both public and private, to work on the completion of the human genome sequence, and afterwards in the analysis and application of potential uses for the resulting genome.

References

[1] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860-921 (2001)

[2] Venter, J.C. et al, The sequence of the human genome. Science 291, 1304-1351 (2001)

[3] Aach, J. et al, Computational comparison of two draft sequences of the human genome. Nature 409, 856-859 (2001)

[4] Bork, P. & Copley, R., Filling in the Gaps. Nature 409, 818-820 (2001)

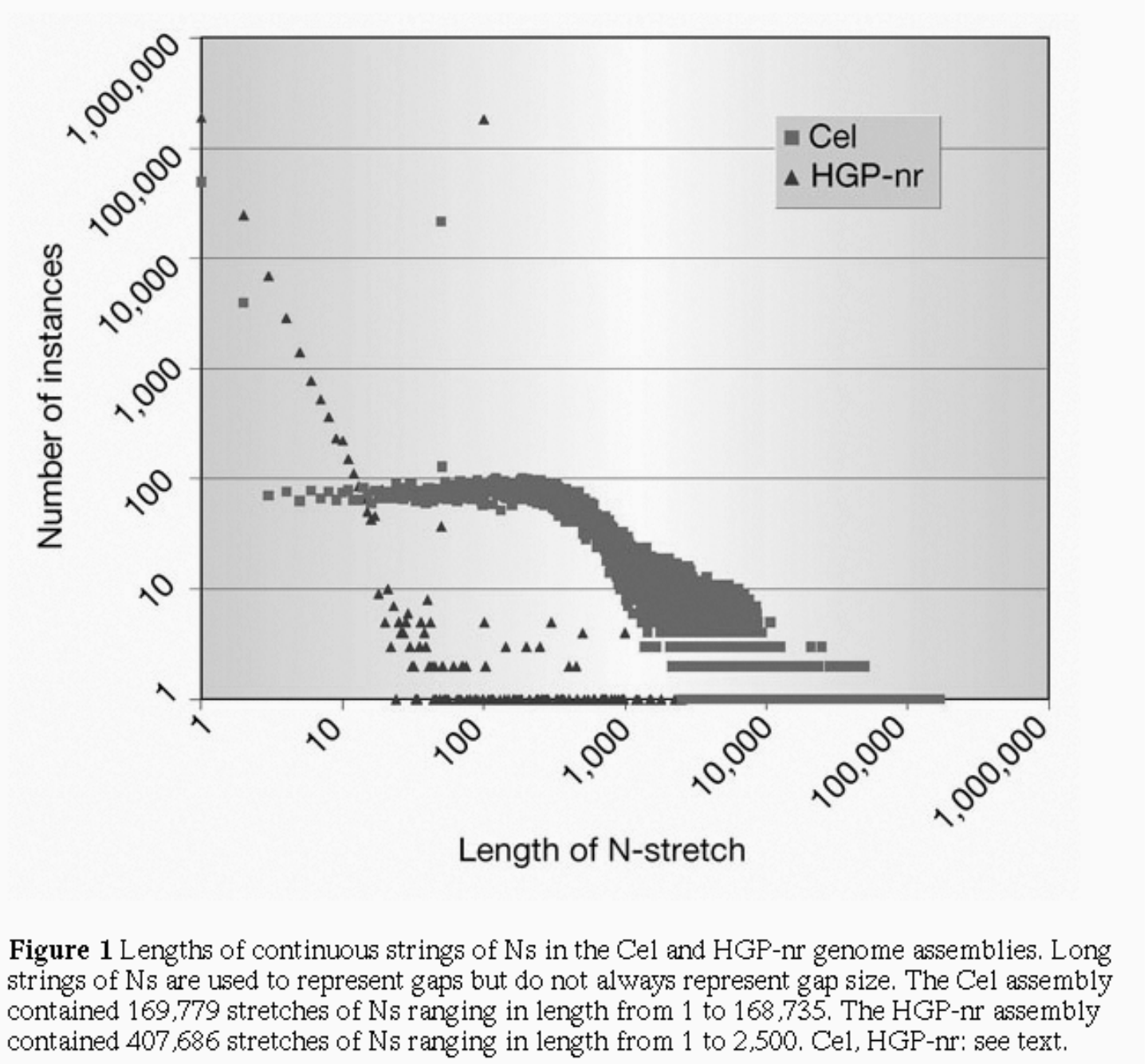

Figure 1 Lengths of continuous strings of Ns in the Cel and HGP-nr genome assemblies. Long strings of Ns are used to represent gaps but do not always represent gap size. The Cel assembly contained 169,779 stretches of Ns ranging in length from 1 to 168,735. The HGP-nr assembly contained 407,686 stretches of Ns ranging in length from 1 to 2,500. Cel, HGP-nr: see text.

Figure 1: Reproduced from Aach et al. [3]