Deposited research article

Assembling and gap filling of unordered genome sequences through gene checking

Amit K. Maiti*, Patrice Bouvagnet

Address: Laboratoire de Génétique Moléculaire Humaine, Université Claude Bernard Lyon 1, 8 av Rockefeller, F-Lyon cedex 08, France. *Present address & Correspondence. Division of Medical Genetics University of Geneva Medical School, 1, rue Michel Servet CH-1211 Geneva , Switzerland

Correspondence: Amit K Maiti. E-mail: Amit.Maiti@medecine.unige.ch, amit@dplanet.ch

Posted: 7August 2001

Genome ***Biology*** 2001, **2(9)**:preprint0008.1-0008.11

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/9/preprint/0008

Received: 2 August 2001

This is the second version of this article to be made available publicly; an earlier version included A. Chakravarti as an autor.
This article was also submitted to *Genome* ***Biology*** for peer review.

Assembling and gap filling of unordered genome sequences through gene checking

Amit K. Maiti*, Patrice Bouvagnet

Laboratoire de Génétique Moléculaire Humaine, Université Claude Bernard Lyon 1,

8 av Rockefeller, F-Lyon cedex 08, France.

*Present address & Correspondence.
Division of Medical Genetics
University of Geneva Medical School,
1, rue Michel Servet CH-1211
Geneva , Switzerland

E.mail. Amit.Maiti@medecine.unige.ch
amit@dplanet.ch

BACKGROUND

The first draft of human genome sequencing is complete. A large amount of DNA sequences are already available in the database but these are not ordered and assembled. In many cases, these sequences are shorter sequences (ranging from 10kb to 100kb) and are separated by "NNNNNN". Also a considerable amount of gaps are to be filled in the subsequent years. Even after generating raw data, properly ordered, finished available sequences, are enormous tasks and expected to take another 2 years.

RESULTS

Here, we describe a simple way to order random genome sequences and to trace gaps. These gaps could be filled by subsequent hybridizations and sequencing. These could be achieved by a simple method by three steps. 1) Selection of large cDNAs in the database (from lower organisms to human). 2) Blasting with these large cDNAs to the unordered human genomic sequences (raw BAC DNA sequences or large DNA fragments) . 3) Ordering these BACs DNA sequences or large DNA fragments based on the homology with cDNA sequences to maintain the continuity of exonic sequences. Homologous exons could also be taken into account on the basis of evolutionary conservacy when other organism's sequence except human, would be used for blasting. Any discontinuity in the exonic sequences denote possible gaps in between two BACs or two sequences.

CONCLUSION

In this way a large number of BACs could be arranged. Subsequently gaps could be traced and filled by further hybridizations and sequencing.

BACKGROUND

Human genome sequencing is in its second phase. Two smallest chromosome 21, 22 and the first draft of whole human genome sequences are already finished [1, 2, 3]. Whole genome sequencing of mouse and rats are in progress. It is expected that a number of other organism's (other primates, dog, cats, etc) genome sequences will be initiated. A large amount of human genome sequences are already available in the database but these sequences are not ordered and assembled. In many cases, these sequences are shorter sequences (ranging from 10kb to 100kb) and are separated by "NNNNNN". From this stage of sequencing efforts, complete and matured finished sequences, ideally without any gaps, are expected. Two principle methods of genome sequencing are currently being employed -systemic sequencing of chromosomal integrated BACS [4] and shotgun sequencing followed by assembling of these sequences in right order into the chromosome [5]. By systematic sequencing approach, ideally all BACs should be ordered and integrated. Given the enormous tasks for a genome sequencing, a large number of BACs are not ordered and overlapped, thus gaps in finished sequences remain. By shotgun sequencing method, ordering and assembling these sequences are the only concern. Even after generating raw data (after completion of first draft) by both of these methods, properly ordered, finished available sequences are expected to take another 2 years [6]. The ordering of nonoverlapping BACs sequences and assembling these sequences into a whole chromosome are a major concern [7]. A "BAC selection parking

strategy" was suggested to minimize the cost and to order BACs after sequencing [8]. This employs minimal 10% overlapping sequences in the BAC ends and subsequent walking. Moreover, tracing the extent of gaps and filling of these gaps are important to integrate these sequences into a chromosome. We propose a gene hunting based method for ordering these BACs sequences, tracing and filling these gaps to obtain a proper integration of these sequences into chromosomes.

RESULTS

This could be achieved by a simple method by three steps. 1) Selection of large cDNAs in the database (from lower organisms to human). There are no shortages of long cDNAs and plenty of (more than 83000 complete cDNAs from all organisms are already in the GeneBank) are being accumulated. A large number of more than 5kb long cDNAs ((only KIAA [9] group represents more than 100 cDNA over 4kb) are in the database. 2) Blasting with these large cDNAs to the unordered human (or any other organism's) raw genomic sequences. These blasts give homology with corresponding exonic sequences. 3) Since blasts are to be done with long cDNA or complete gene, two or more DNA sequences (coming from BACs) could be ordered based on continuity of exonic sequences. A few examples of proper ordering are shown in table 1. The corresponding cDNAs are blasted with human "high throughput genome sequences" (htgs). BAC sequences are ordered on the basis of evolutionary conservacy of the exonic sequences.

The most important information about the position and the extent of gap could be obtained from the ordered BACs. In the above ordering (table1), exonic gaps (missing

exon(s)s or the part of exons which does not fit with the continuous cDNA sequences; denoted by * in the table 1) in KIAA0535 (650-720), RATMIBP1 (368-615) and in RalGPS1A (484-565, 875-1015) are 70bp, 247bp, 81bp and 140bp respectively. Even if large intron(s) lying in between two consecutive BACs in these cases, it is not likely that the intron size could be more than 50kb sequences. Unknown BACs could be traced from BAC library by hybridization with corresponding partial cDNA and further sequencing of the selected BACs could fill gaps.

These blasts could also be informative to order a number of BACs sequences carrying homologous genes. When blasted with single cDNA, mouse lrd (left right dynein) gene arranges BACs sequences carrying two different homologous dynein heavy chain genes in chromosome 17 (table 2). Although, the power of the identity/similarity decreases due to decreasing homology with the lrd but these should not interfere to deduce the exonic continuity and subsequently ordering of consecutive BACs carrying homologous genes.

These blasts have been done with the available genomic DNA (BACs sequences) deposited from all the chromosomes (htgs). When these blasts could be done with BACs DNA sequences from single chromosome or part of a chromosome that most genome center does, results should be robust. Over 30000 [10] genes are already assigned into chromosome by radiation hybrid and assignments of more genes into chromosomes are being rapidly accumulated. A large number of complete cDNAs have been characterized and are being characterized. These mapped genes are immensely helpful for chromosome

specific blasting. Those shorter sequences (less than 100kb) arising from sequencing initially could be ordered by this way and finally assembled by simple blast or by "BAC end Power Blast"[7]

CONCLUSION

By this way all unordered BACs cannot be ordered since it depends on continuous or overlapping exonic sequences. Nevertheless, this is an efficient way to order a large number of genes carrying BACs sequences. Chromosome 21 and chromosome 22 have 10.5Mb and 3Mb sequences geneless [1,2] region. Even if one third of the human genome is geneless (but that is not the actual case [3]), total coding region spans through 2 billion of 3 billion bp and lies within 10000 to 15000 BACs (average BAC size is 150kb to 200kb) approximately. Also blasting with large cDNA would be useful when difficulties or ambiguities arise during arranging two or more BAC DNA sequences of unknown location. Apart from human, a many organism's genome sequencing has been initiated and draft sequence is being generated. A hybrid procedure of both systematic sequencing [4] and shotgun sequencing [5] has been recently conceded for sequencing other genomes. Complete or partial genes (long cDNAs but not ESTs) could be useful to order two sequences or trace gaps. It i s one of a simple way to integrate BACs sequences into chromosomes or to order random short sequences and to obtain an accurate, properly finished sequence of each chromosome.

ACKNOWLEDGEMENT

We thank Prof. Stylianos Antonarakis for reading and critical comments on this manuscript.

REFERENCES

1. Dunham, I. *et al*. **The DNA sequence of human chromosome 22.** *Nature* **402**, 489-495 (1999).
2. Hattori, M. *et al*. **The DNA sequence of human chromosome 21** .*Nature* **405**, 311-319 (2000).
3. Lander et al. **Initial sequencing and analysis of the human genome.** *Nature* **409**, 860 - 921 (2001)
4. Marshal, E A. **A strategy for sequencing the genome 5 years early** . *Science* **267,** 783-784 (1995).
5. Fleishmann, RD. *et al*. **Whole genome random sequencing & assembly of Haemophylus influenzae.** *Science* **269**, 496-511 (1995).
6. Pennisi, E. **Finally the book of life and instructions for navigating it** . *Science* **288**, 2304-2307 (2000)

7. Kuehl, PM., Weisemann, JM., Touch man, JW., Green, ED., Boguski, MS. **An effective approach for analyzing "prefinished" genomic sequence data** . *Genome Res.* **9(2)**, 189-94 (1999).

8. Roach, JC., Siegel, AF., Ench, GD., Trask, B and Hood, L. **Gaps in the Human Genome Project.** *Nature* **401,** 843-845 (1999)

9. Nagase, T. *et al*. **Prediction of the coding sequences of unidentified human genes. XX. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro**. *DNA Res*. **4 (2)**, 141-150 (1997).

10. Deloukas, P. *et al*. **A physical map of 30,000 human genes.** *Science*. **282**, 744-746 (1998)

Table1: Ordering of BACs sequences by blasting with long cDNAs to the human genome sequences (htgs). Exonic gaps and subsequently missing BACs are denoted by star (*). In cases of missing BACs sequences, orderings are denoted as (1),(2),(*4) instead of (1), (2), (3).

cDNAs & Acc. No	Length (bp)	Exon sequences homologous to BACs sequences (bp)	Acc. No., length (bp) & ordering of BACs (denoted by first as 1, Second as 2, and so on)	Chromosome[$]
Sea Urchin ß-DYNEIN (X59603)	13799	(2143-9465) (9228-12323)	AC073102.1, 197994bp (1) AC004002.1, 112314bp (2)	7
RERE protein (AB036737)	8035	(3-492) (491-1467) (1158-1921) (1834-8018)	AC025240.3, 170048bp (1) AC021953.3, 169752bp (2) AL096855.26,177246bp (3) AC016049.2, 198254bp (4)	1
KIAA0603 (NM_014832)	5922	(118-2380) (842-5922) (1516-5922)	AL139230.7, 217520bp (1) AC011164.4, 183577bp (2) AL162571.4, 184953bp (3)	13
KIAA0570 (NM_014709)	5842	(64-3895) (120-4205) (2576-5843)	AC073568.1, 138091bp (1) AC016894.4, 164561bp (2) AC018889.2, 174186bp (3)	2
KIAA0535 (NM_014682)	6183	(1-720) (650-3047)* (3129-6183)	AC009995.4, 158883bp(1) AC021915.1, 51198bp (2) AC023536.1, 50388bp (*4)	8
RalGPS1A (NM_014636)	6336	(1-484)* (565-875)* (1015-6336)	AL160169.2, 145953bp(1) AL356862.2, 132794bp(*3) AL354705.3, 165901bp(*5)	9
KIAA0294 (NM_014629)	8467	(1-5983) (5206-8467)	AF236876.2, 113733bp(1) AF170801.2, 131032bp(2)	8
(RATMIBP1) (D37951)	9731	(1-368)* (615-9731)	AL355307.2, 206594bp(1) AL136122.3, 131255bp(*3)	6

[$] Chromosome assignments are according to release of BAC sequences from genome centers.

.

Table 2: Ordering of 2 homologous gene carrying BACs by blasting with a single gene, Mouse lrd.

cDNA & Acc. No	Length (bp)	Exonic sequences homologous to BAC sequences (bp)	Acc. No., length (bp) & ordered BACs	Chromosome[$]
Mouse lrd (AF183144)	14088	(5459-7701) (4015-12515)	AC023413, 132649bp (1) AC016182, 173180bp (2)	17
Mouse lrd (AF183144)	14088	(3637-7701) (7709-12714) (13189-13602)	AC005701, 197455bp (1) AC005209, 184130bp (2) AC005410, 139049bp (3)	17

[$] Chromosome assignments are according to release of BAC sequences from genome centers.