# CHAPTER 6

# Sequence Analysis of the Evolutionary Rearrangement Points in the Gibbon

## 6.1 Introduction

The work described so far in this thesis has involved the mapping of evolutionary rearrangement points in gibbon chromosome 18 at increasing levels of resolution. The boundaries of the human chromosome 22-homologous blocks were defined in gibbon chromosome 18 by cross-species chromosome painting to a resolution of approximately 7 Mb. The identification of a human chromosome 22 fosmid clone hybridising to both arms of gibbon chromosome 18 increased the resolution of the analysis of one of the rearrangement points to within 40 kb of sequence. The rearrangement point was further narrowed down to within 1 kb of sequence by STS PCR carried out on gibbon cosmid clones spanning the fusion points between human chromosome 22- and human chromosome 16-homologous blocks on gibbon 18 p and q.

In order to analyse the fusion regions at the highest resolution possible, it was necessary to generate sequence from the gibbon cosmid clones isolated in chapter 5. The initial strategy used was to amplify fragments containing gibbon sequences spanning the fusion points by vectorette PCR. These products would then be used as sequencing templates. This approach was taken because of the advantages of the vectorette system in terms of speed and resources: it was an economical and rapid method to confirm that the cosmids do contain sequences spanning the fusion points, perhaps before embarking on the more costly and time-consuming sequencing of two entire cosmids. Vectorette PCR was originally developed by Riley et al., (1990) for the rescue of the ends of YAC, and it is an efficient method for isolating unknown DNA adjacent to any known sequence of 20 bp or more in length. However, there are potential drawbacks to using this method as it generates short sequences and, thus, provides only a limited analysis of the region of interest. For that reason, the sequencing of the entire cosmids was planned as an alternative strategy.

## RESULTS

## 6.2 Isolating gibbon fusion point fragments for sequence analysis

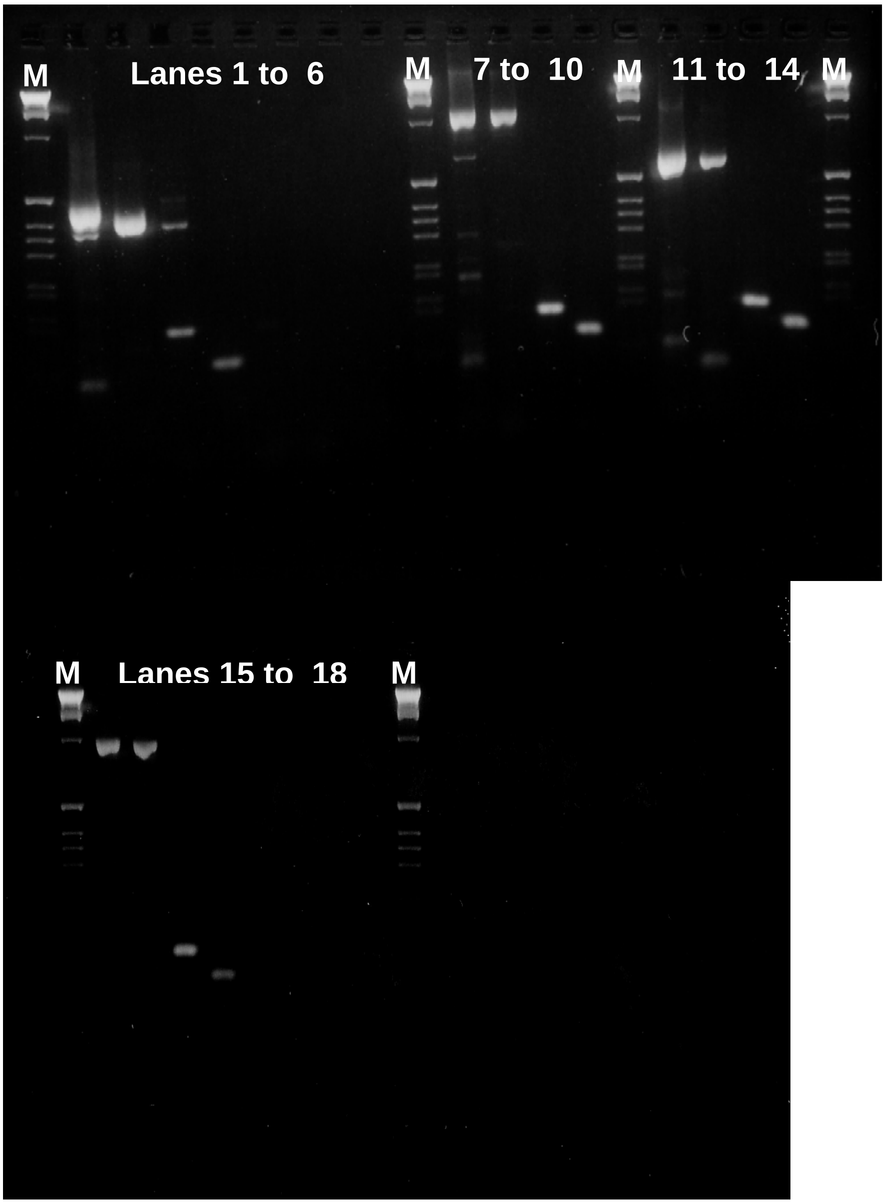
### 6.2.1 <u>Breakpoint fragment isolation by vectorette PCR</u>

According to STS mapping (Chapter 5), the sequence on human chromosome 22 corresponding to the gibbon ancestral chromosome 18 rearrangement breakpoint could be narrowed down to a 1 kb stretch of DNA between STSs B1 and B2. In order to provide additional information about the sequences spanning the fusion points on gibbon chromosome 18, fragments of SCHSY4.11 and SCHSY4.21 were isolated using a vectorette system modified to use human primers to carry out PCR on gibbon cosmids. For SCHSY4.11 the B1 sense primer was used with the vectorette primer 224. For SCHSY4.21 the B2 antisense was used with 224. A single product of approximately 400 bp was generated from SCHSY4.11 digested with *RsaI*. For SCHSY4.21 single products were generated following digestion with *Pst*I*, Rsa*I and *Hinc*II of sizes 1.0 kb, 550 bp and 900 bp, respectively (figure 6.1). The identity of the four PCR products was confirmed by hemi-nested PCR, which also served to generate large amounts of DNA for sequencing. For the SCHSY4.11 *Rsa*I B1 + 224 product a nested primer “B1 nest” was designed (Sarah Hunt) from the human sequence distal to B1. For the three SCHSY4.21 B2 + 224 products, a nested primer “B2 nest” was designed proximal to B2. Primer 224 was used in combination with B1 nest or B2 nest for these secondary PCRs (see figure 6.1).

The vectorette PCR products from SCHSY4.11 *Rsa*I B1 + 224 (approximately 400 bp) and SCHSY4.21 *Pst*I B2 + 224 (approximately 1.2 kb) were electrophoresed and visualised on a preparative agarose gel. The appropriate bands were excised and the DNA extracted from the gel using the Geneclean ™ kit.

*Figure 6.1* (next page) Agarose gel analysis of vectorette PCR products. M= 1 kb DNA size marker, sizes as indicated. Lanes as follows:

1. SCHSY4.11 *Rsa*I product from B1sense + 224
2. SCHSY4.11 *Rsa*I product from B1nest + 224
3. SCHSY4.11 *Rsa*I product from B1sense + B1antisense
4. SCHSY4.11 *Rsa*I product from B1nest + B1antisense
5. Gibbon genomic DNA control product from B1sense + B1antisense
6. Negative control
7. SCHSY4.21 *Pst*I product from B2antisense +224
8. SCHSY4.21 *Pst*I product from B2nest + 224
9. SCHSY4.21 *Pst*I product from B2antisense + B2sense
10. SCHSY4.21 *Pst*I product from B2nest + B2sense
11. SCHSY4.21 *Rsa*I product from B2antisense + 224
12. SCHSY4.21 *Rsa*I product from B2nest + 224
13. SCHSY4.21 *Rsa*I product from B2antisense + B2sense
14. SCHSY4.21 *Rsa*I product from B2nest + B2sense
15. SCHSY4.21 *Hinc*II product from B2antisense + 224
16. SCHSY4.21 *Hinc*II product from B2nest + 224
17. SCHSY4.21 *Hinc*II product from B2antisense + B2sense
18. SCHSY4.21 *Hinc*II product from B2nest + B2sense


M
Lanes 1 to 6
M
7 to 10
M
11 to 14
M
M
Lanes 15 to 18
M

### 6.2.2 Sequence analysis of vectorette PCR fragments

The vectorette-PCR fragments were sequenced by Elizabeth Huckle of the Sanger Institute Sequencing Development Team. Each product was sequenced from both directions using primers 224 and B1 nest for SCHSY4.11 or 224 and B2 nest for SCHSY4.21.

The SCHSY4.21 reaction primed by B2 nest yielded a stretch of 508 bases of sequence homologous to human chromosome 22 distal to the B2 STS. The SCHSY4.21 reaction primed by 224 yielded a stretch of 460 bp of non-human homology (closest to the priming site) followed by a 132 bp stretch of sequence homologous to human chromosome 22.

The SCHSY4.11 sequencing reaction primed by B1 nest yielded a stretch of 243 bp of sequence homologous to human chromosome 22 proximal to the B1 STS. The SCHSY4.11 reaction primed by 224 yielded a stretch of 111 bases of sequence homologous to human chromosome 22 proximal to the B1-primed region. Preceding this sequence, there were a further 15 bases of sequence which did not match any known human sequence by BLAST.

It was possible that the un-matched sequences from both vectorette products were actually homologous to unsequenced regions of human chromosome 16 (or another human chromosome), or that they represented chromosomal material present in the great ape ancestral genome, which was lost since divergence of the lesser apes occurred. However, at this stage the possibility could not be excluded that there are gibbon-specific sequences flanking chromosome 22 homologous material on both sides of the fusion points. The gibbon sequence may have been inserted at the rearrangement points at the time of or after the rearrangement event.

The last seven bases before the homology with human chromosome 22 is lost are common to SCHSY4.21 and SCHSY4.11. The region of commonality and loss of homology was in an Alu repeat.

From the sequence generated from the vectorette PCR products, it was difficult to speculate about the rearrangement mechanism. As the PCR products generated such short sequencing reads, limitations of the vectorette PCR approach have been demonstrated. For example, if the rearrangement were sponsored by complex sequence motifs farther than a few hundred bp away from the junction points, it would be impossible to explore the mechanism.

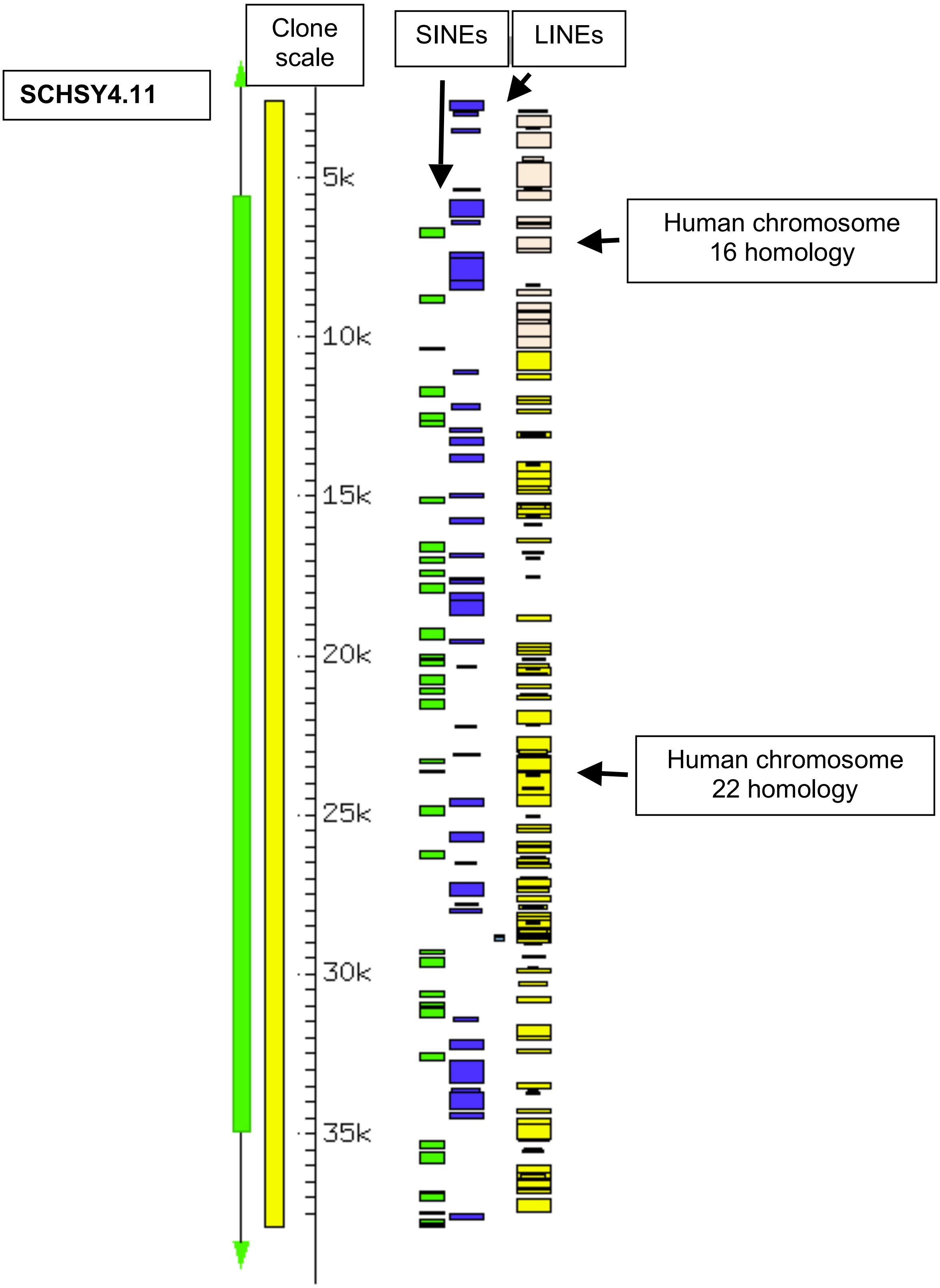
In order to describe the rearrangement points in detail (and generate further evidence for a possible mechanism) a full sequence analysis of clones SCHSY4.11 and SCHSY4.21 was carried out.

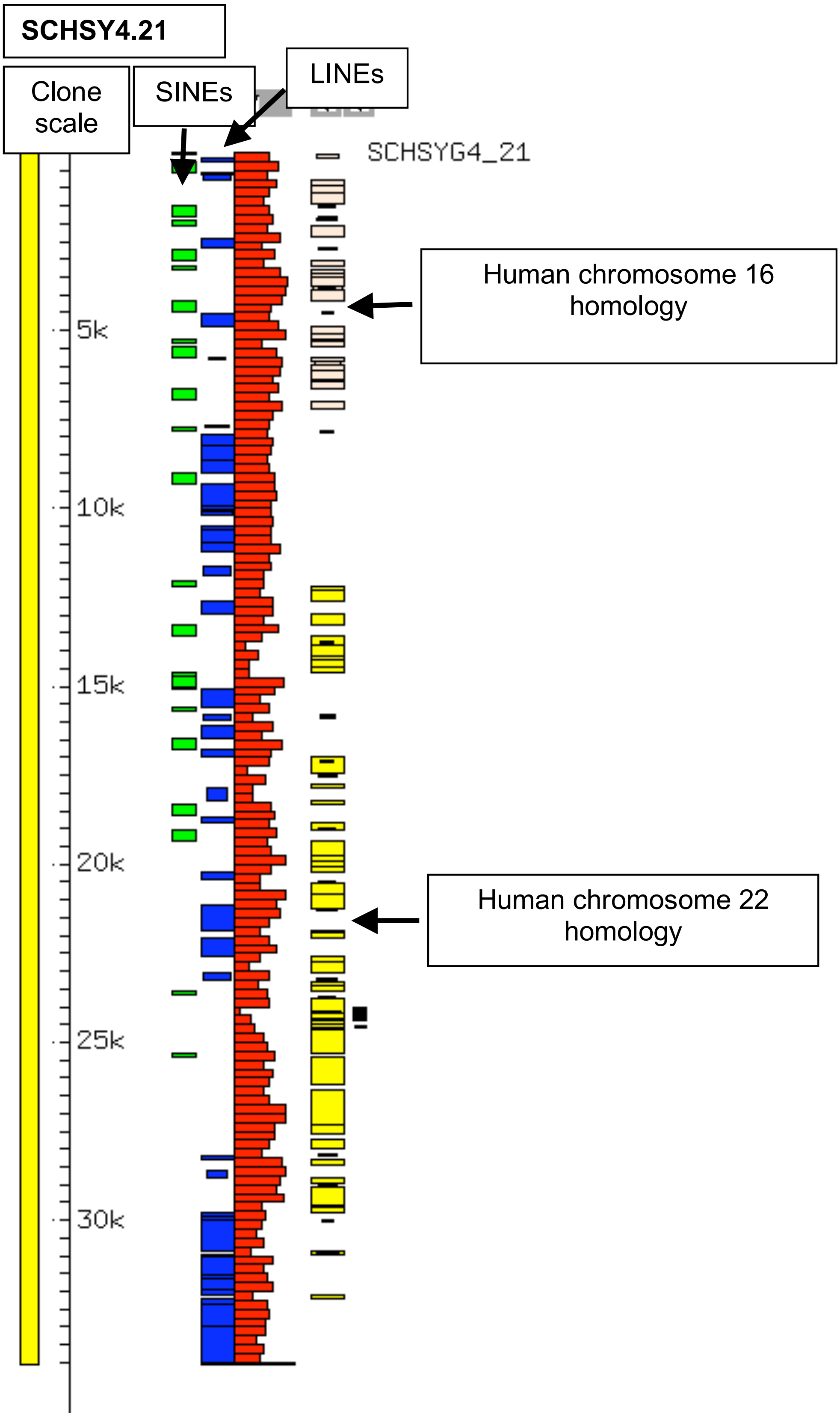
## 6.3 Sequence analysis of SCHSY4.11 and SCHSY4.21 cosmids

DNA from the two cosmids was prepared using a standard alkaline lysis procedure and was submitted to the Sanger Centre sequencing pipeline (overseen by Matt Jones, David Willey, and Kirsten McClay). Highly accurate, finished sequence was generated for each clone, and the sequences were submitted to the EMBL database. The sequence data were analysed for repeat elements (RepeatMasker), DNA homologies (BLASTN) and gene predictions (FgenesH) by Sarah Hunt (Sanger Institute Informatics Group). Along with other annotation information, the results of the analysis were stored in Gibbonace (created by Carol Scott and Sarah Hunt), which is an implementation of ACeDB.

The DNA insert of clone SCHSY4.11 is 40,516 bp long and the insert of clone SCHSY4.21 is 34,056 bp long. The sequences from both junction point cosmids support the lower-resolution analyses, described previously, which placed the HSA 16 and HSA 22 homologous blocks adjacent to each other on either side of the rearrangement junctions on HSY 18p and 18q. The gross features of each clone are summarised in Figure 6.2 and 6.3.

*Figure 6.2 and 6.3* (next two pages) Screen capture from Gibbonace illustrating major sequence landmarks of clones SCHSY4.11 and 4.21.


SCHSY4.11
Clone scale
SINEs
LINEs
5k
10k
15k
20k
25k
30k
35k
Human chromosome 16 homology
Human chromosome 22 homology


SCHSY4.21
Clone scale
SINEs
LINEs
SCHSYG4_21
5k
10k
15k
20k
25k
30k
Human chromosome 16 homology
Human chromosome 22 homology

### 6.2.2 Sequence analysis of vectorette PCR fragments

The vectorette-PCR fragments were sequenced by Elizabeth Huckle of the Sanger Institute Sequencing Development Team. Each product was sequenced from both directions using primers 224 and B1 nest for SCHSY4.11 or 224 and B2 nest for SCHSY4.21.

The SCHSY4.21 reaction primed by B2 nest yielded a stretch of 508 bases of sequence homologous to human chromosome 22 distal to the B2 STS. The SCHSY4.21 reaction primed by 224 yielded a stretch of 460 bp of non-human homology (closest to the priming site) followed by a 132 bp stretch of sequence homologous to human chromosome 22.

The SCHSY4.11 sequencing reaction primed by B1 nest yielded a stretch of 243 bp of sequence homologous to human chromosome 22 proximal to the B1 STS. The SCHSY4.11 reaction primed by 224 yielded a stretch of 111 bases of sequence homologous to human chromosome 22 proximal to the B1-primed region. Preceding this sequence, there were a further 15 bases of sequence which did not match any known human sequence by BLAST.

It was possible that the un-matched sequences from both vectorette products were actually homologous to unsequenced regions of human chromosome 16 (or another human chromosome), or that they represented chromosomal material present in the great ape ancestral genome, which was lost since divergence of the lesser apes occurred. However, at this stage the possibility could not be excluded that there are gibbon-specific sequences flanking chromosome 22 homologous material on both sides of the fusion points. The gibbon sequence may have been inserted at the rearrangement points at the time of or after the rearrangement event.

The last seven bases before the homology with human chromosome 22 is lost are common to SCHSY4.21 and SCHSY4.11. The region of commonality and loss of homology was in an Alu repeat.

From the sequence generated from the vectorette PCR products, it was difficult to speculate about the rearrangement mechanism. As the PCR products generated such short sequencing reads, limitations of the vectorette PCR approach have been demonstrated. For example, if the rearrangement were sponsored by complex sequence motifs farther than a few hundred bp away from the junction points, it would be impossible to explore the mechanism.

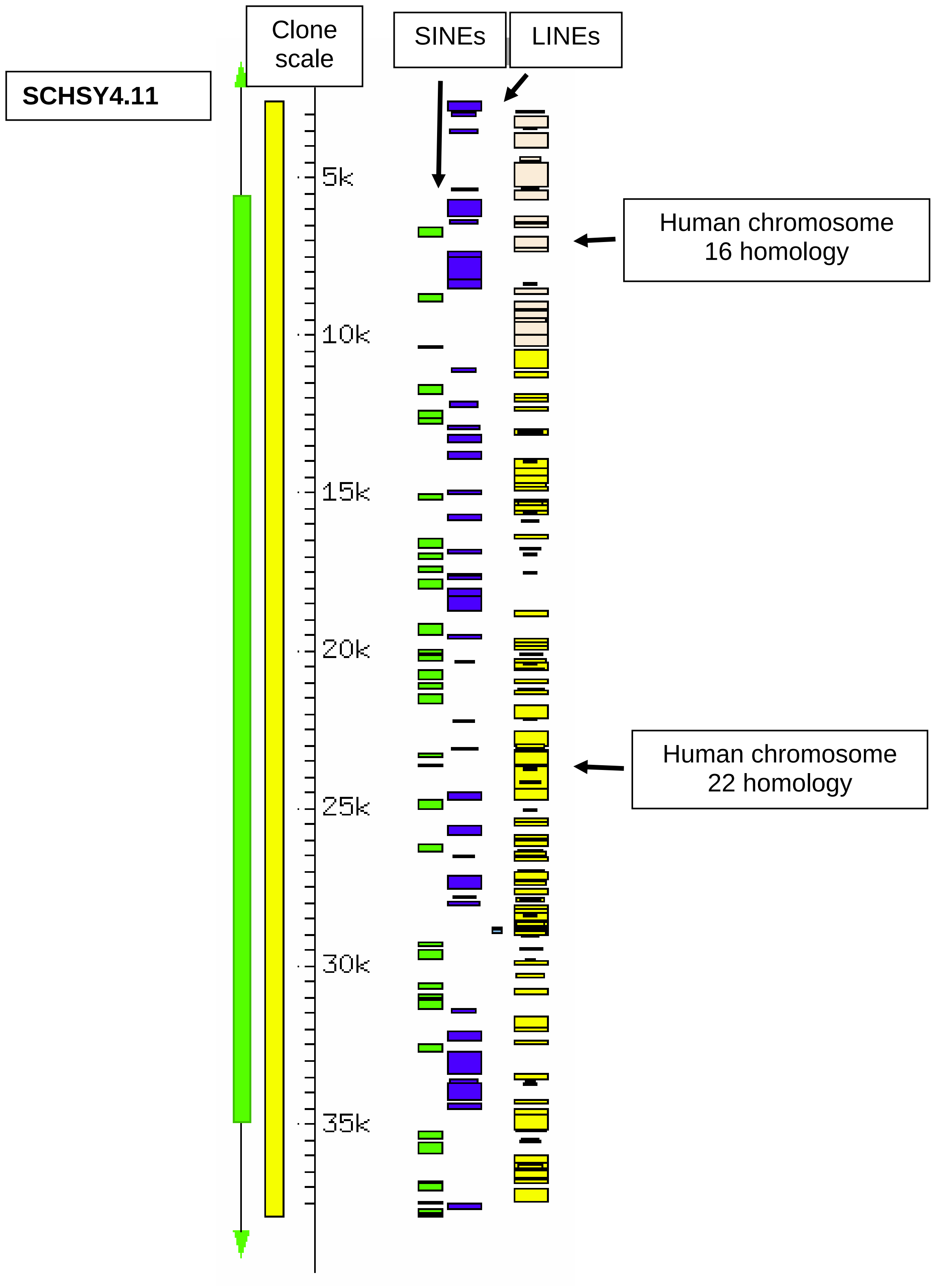
In order to describe the rearrangement points in detail (and generate further evidence for a possible mechanism) a full sequence analysis of clones SCHSY4.11 and SCHSY4.21 was carried out.

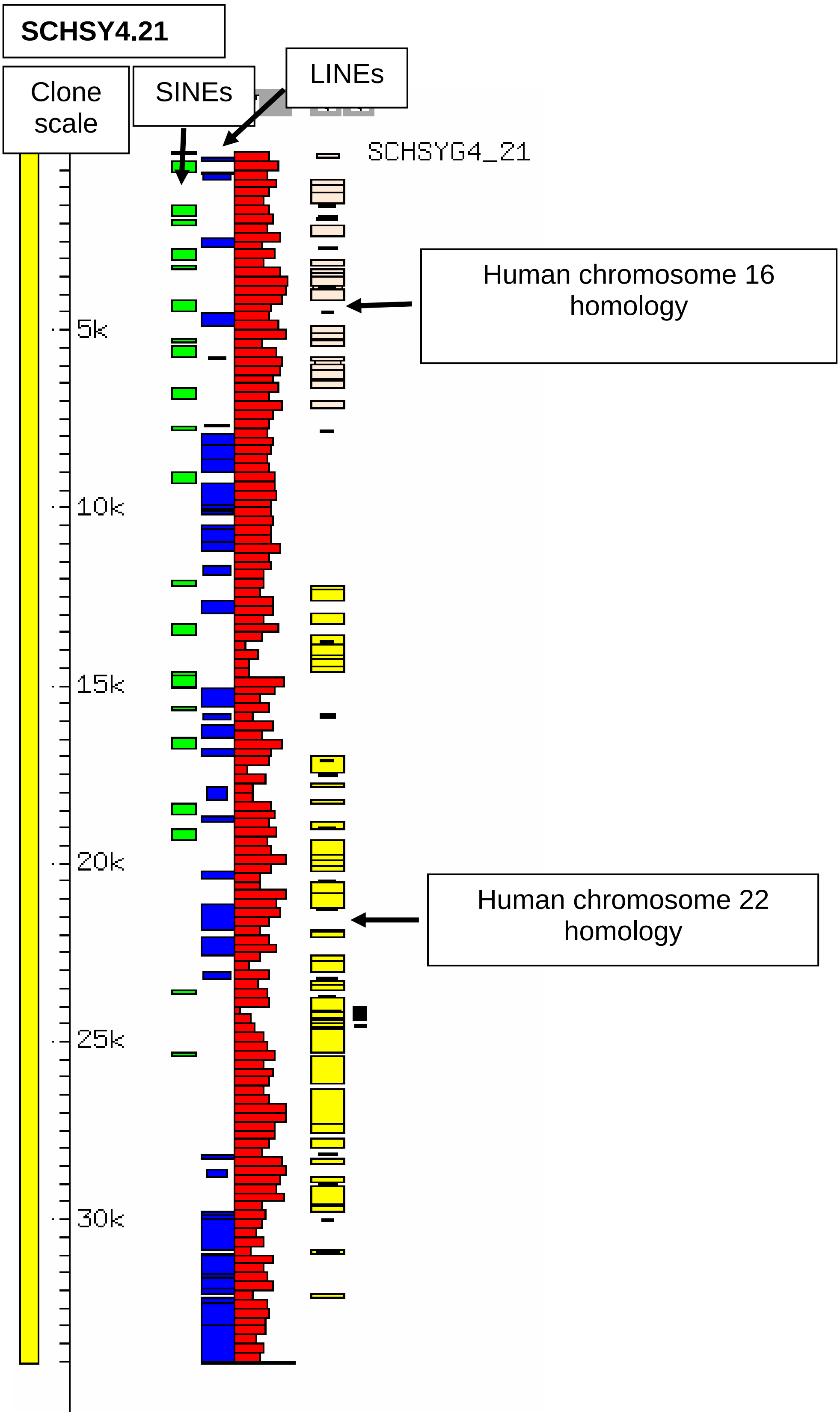
### 6.3 Sequence analysis of SCHSY4.11 and SCHSY4.21 cosmids

DNA from the two cosmids was prepared using a standard alkaline lysis procedure and was submitted to the Sanger Centre sequencing pipeline (overseen by Matt Jones, David Willey, and Kirsten McClay). Highly accurate, finished sequence was generated for each clone, and the sequences were submitted to the EMBL database. The sequence data were analysed for repeat elements (RepeatMasker), DNA homologies (BLASTN) and gene predictions (FgenesH) by Sarah Hunt (Sanger Institute Informatics Group). Along with other annotation information, the results of the analysis were stored in Gibbonace (created by Carol Scott and Sarah Hunt), which is an implementation of ACeDB.

The DNA insert of clone SCHSY4.11 is 40,516 bp long and the insert of clone SCHSY4.21 is 34,056 bp long. The sequences from both junction point cosmids support the lower-resolution analyses, described previously, which placed the HSA 16 and HSA 22 homologous blocks adjacent to each other on either side of the rearrangement junctions on HSY 18p and 18q. The gross features of each clone are summarised in Figure 6.2 and 6.3.

*Figure 6.2 and 6.3* (next two pages) Screen capture from Gibbonace illustrating major sequence landmarks of clones SCHSY4.11 and 4.21.


SCHSY4.11
Clone scale
SINEs
LINEs
5k
10k
15k
20k
25k
30k
35k
Human chromosome 16 homology
Human chromosome 22 homology


SCHSY4.21
Clone scale
SINEs
LINEs
SCHSYG4_21
5k
10k
15k
20k
25k
30k
Human chromosome 16 homology
Human chromosome 22 homology

SCHSY4.11 has 10.3 kb of homology to human chromosome 16 and the homology immediately switches to human chromosome 22 to the end of the clone (25 kb). The homology switches at the site of a partial AluJo element, which originates from the human chromosome 22 homologous material. The chromosome 16 homology has two locations on human chromosome 16 separated by a distance of 5 Mb.

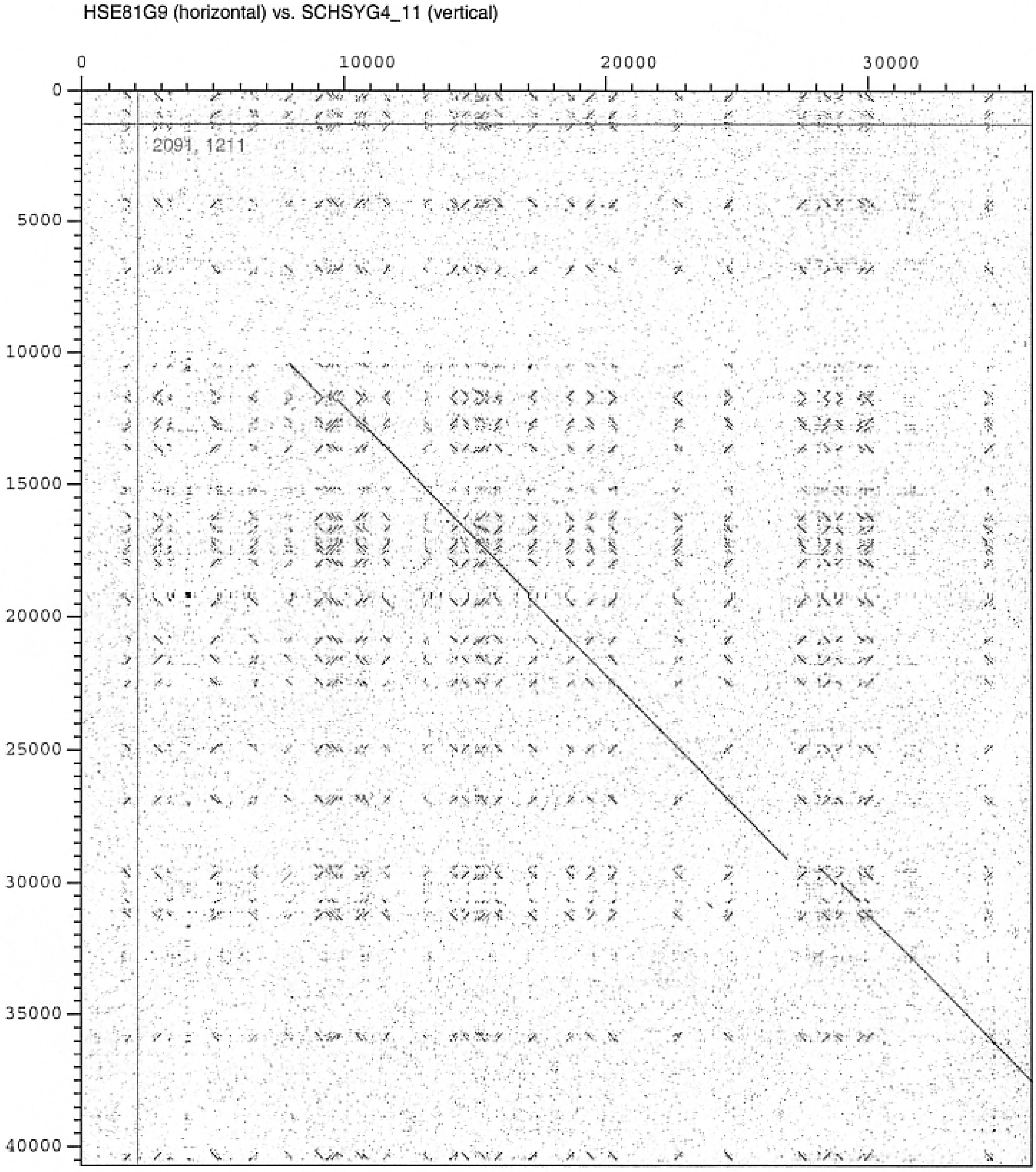
SCHSY4.21 has 7 kb of homology to human chromosome 16, and 25 kb of homology to human chromosome 22 to the end of the clone. The homology to human chromosome 22 starts at the site of a partial AluJo element (bases 21 to 148), which originates from the human chromosome 22 homologous material. There is a stretch of 5 kb of sequence between the chromosome 16 and 22 homologies, which has no human homology. An AluJo element is located at the end of the main part of HSA16 homology, but there is a short section (250 bp) of inverted duplicated HSA16 homologous material after the AluJo.

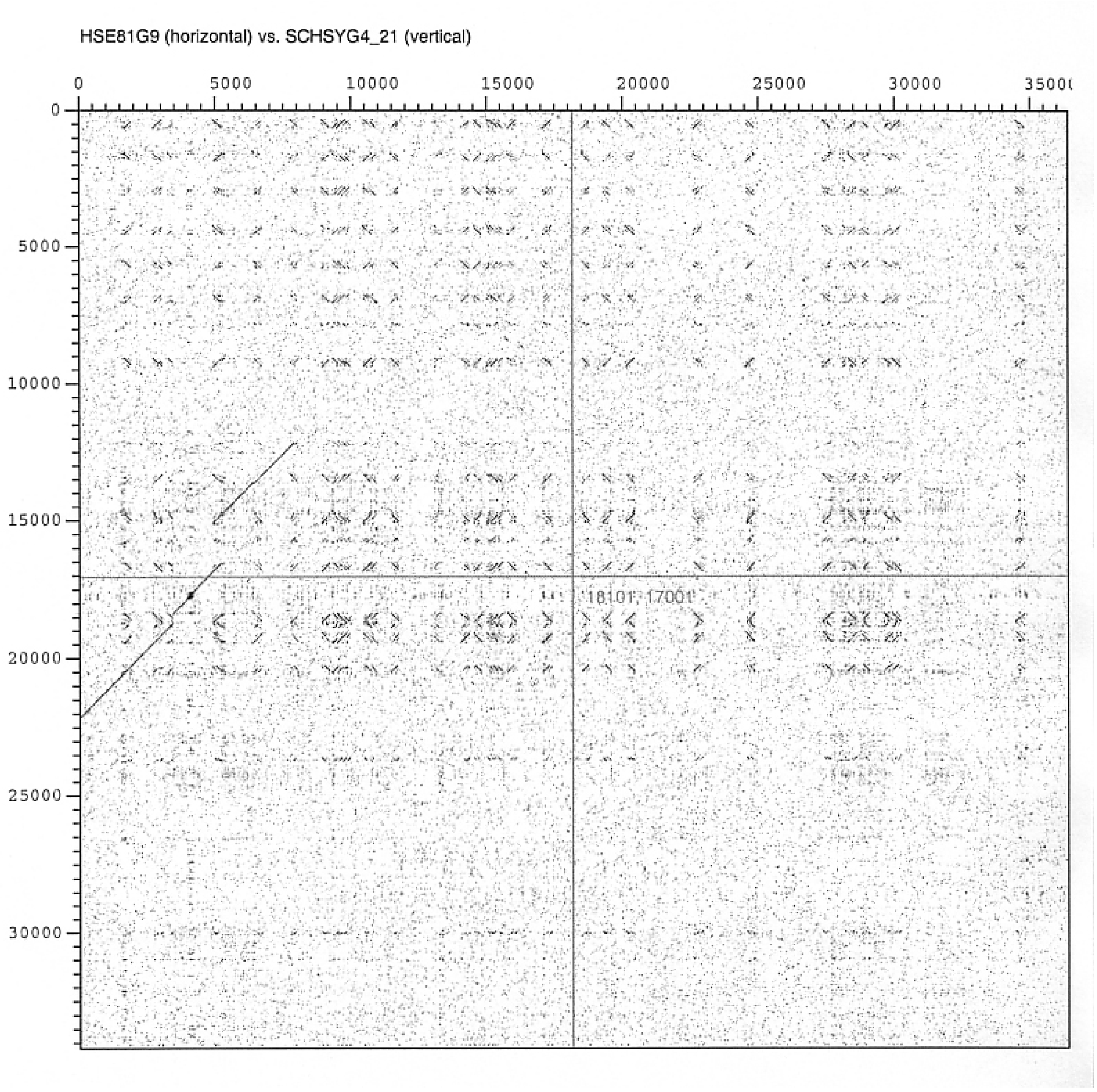
Dotter analyses were carried out to illustrate graphically the homology between the HSA22 breakpoint region (in human chromosome clone HSE81G9), the HSA16 breakpoint region (in sequence AC126763) and the sequences of clones SCHSY4.11 and 4.21.

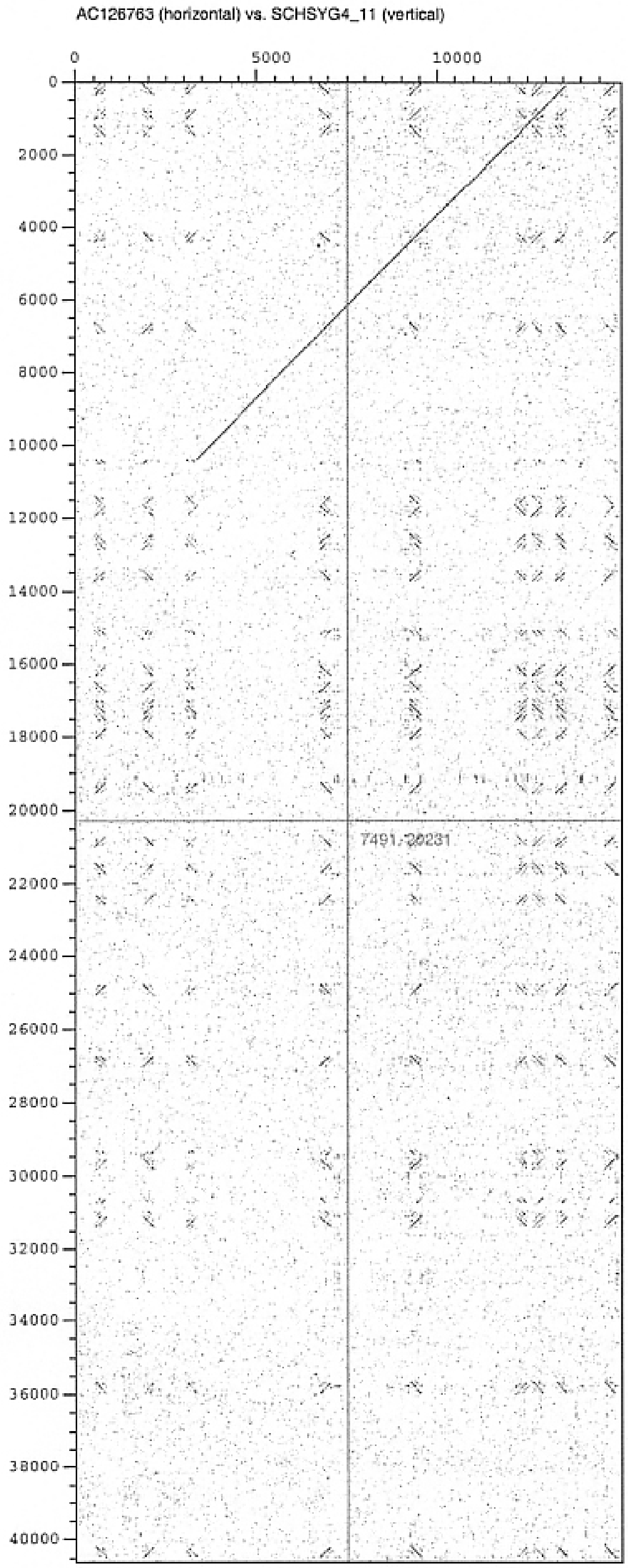
As can be seen in Figure 6.4, there is almost continuous homology between HSE81G9 (from 8 kb to the end) and SCHSY4.11 (from 10 kb to the end). There is also almost continuous homology over 9 kb between the HSA22-homologous section of SCHSY4.21 and HSE81G9 (Figure 6.5). There is one region of 200 bp, which is found in the gibbon sequence and not in the human, illustrated by a gap in the diagonal line.

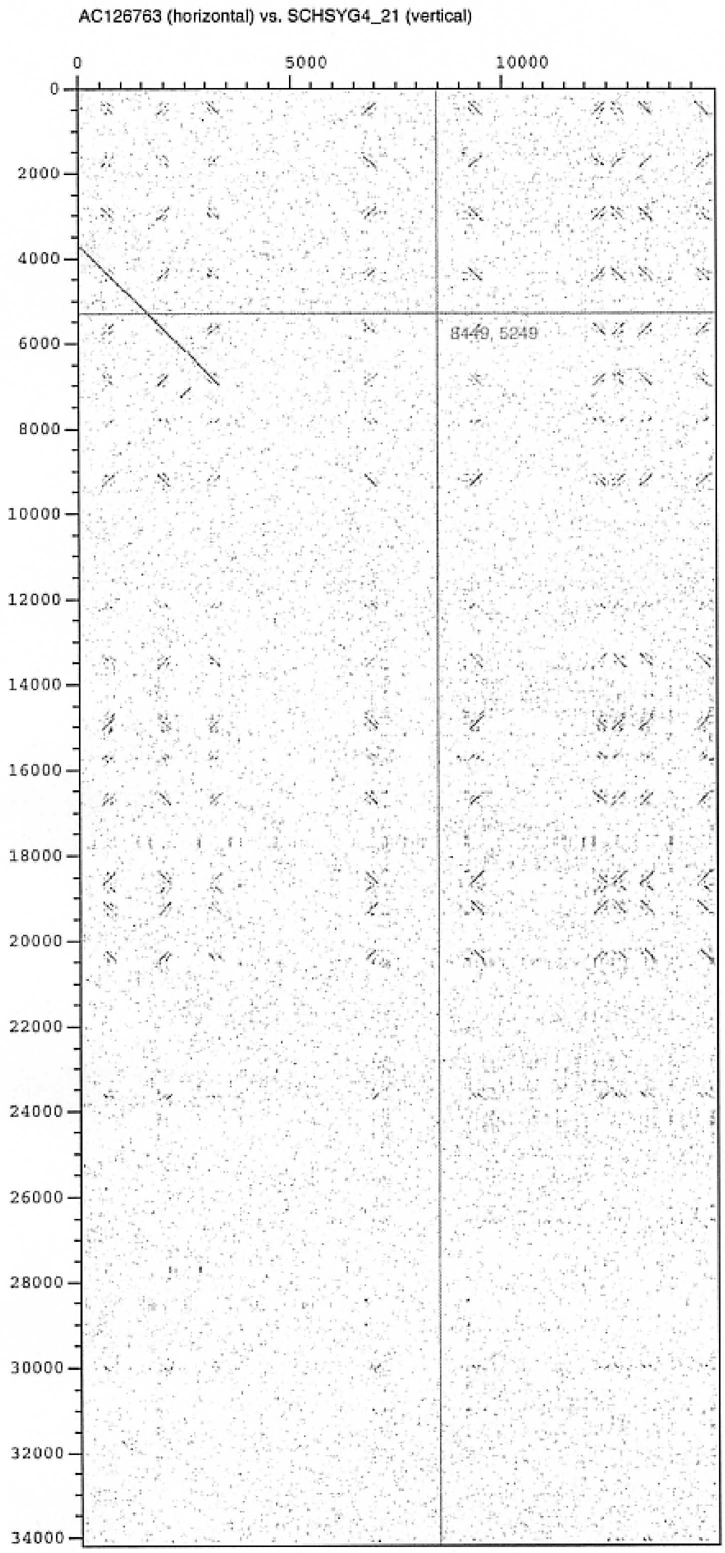
The Dotter output from the comparison of 0-15 kb of AC126763 versus SCHSY4.11 shows continuous homology up to the AluJo element at the breakpoint (Figure 6.6). The analysis from the comparison of AC126763 versus SCHSY4.21shows continuous homology up to the breakpoint (Figure 6.7). After the AluJo in 4.21, there is a small region of inverted homology to an earlier section (2.4-2.6 kb) of AC126763. Other than repeat elements, there is no homology between AC126763 and HSE81G9, or between 4_21 and 4_11.

*Figures 6.4, 6.5, 6.6 and 6.*7 (on the following four pages) illustrate the Dotter analysis outputs from sequence comparisons between HSE81G9 versus SCHSY4.11, HSE81G9 versus SCHSY4.21, AC126763 versus SCHSY4.11 and AC126763 versus SCHSY4.21, respectively.


HSE81G9 (horizontal) vs. SCHSYG4_11 (vertical)
0
10000
20000
30000
0
5000
10000
15000
20000
25000
30000
35000
40000
2091, 1211


HSE81G9 (horizontal) vs. SCHSYG4_21 (vertical)
0
5000
10000
15000
20000
25000
30000
35000
0
5000
10000
15000
20000
25000
30000
18101, 17001


AC126763 (horizontal) vs. SCHSYG4_11 (vertical)
0
5000
10000
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
26000
28000
30000
32000
34000
36000
38000
40000
7491, 20231


AC126763 (horizontal) vs. SCHSYG4_21 (vertical)
0
5000
10000
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
26000
28000
30000
32000
34000
8449, 5249

The partial AluJo elements cloned in 4.11 and 4.21 were blasted against non-redundant human sequence and both were found to be homologous to HSA22. The AluJo element adjacent to the inverted duplicated HSA16 material in 4.21 is homologous to HSA16. The repeatmasker co-ordinates of the partial AluJo elements appear to follow on from each other, that is, the two partial Alu elements represent the same region of an AluJo on the human chromosome 22 sequence. This suggests that the two halves of the AluJo elements (one on 4.11, the other in 4.21) have been derived from a breakage and inversion in human chromosome 22 homologous material.

As a consequence of where the breakage occurred relative to the SINEs, one of the products (4_11) has only 0.5 AluJo and the other (4_21) has 1.5 AluJo elements. This is consistent with an homologous recombination event in which the Alu elements had aligned out of register by one sub-repeat unit. In this situation, each of the elements would be partly derived from HSA22 and partly from HSA16 homologous material. If this model were correct, the unknown 5 kb segment would perhaps have inserted and separated the 1.5 AluJo element into the 1 and the 0.5 seen in 4_21 either side of the 5 kb block.

The model proposed for the rearrangement is of a pericentric inversion of the ancestral chromosome, on which the regions homologous to HSA22 and HSA16 were originally on separate arms. The breakpoint in the HSA22 material was within an AluJo element, and the break in the HSA16 material occurred just after an AluJo element. The incorporation of 5 kb of mostly repetitive sequence and the small duplication of material homologous to HSA16 on one of the inversion products is most likely to have occurred during the rearrangement event. The reasoning for this is that if the section were inserted after the rearrangement event, then the insertion would have had to occur precisely at the point of fusion. Another possibility is that the 5 kb of sequence was in fact part of the ancestral chromosome homologous to human 16, that was lost after the lesser apes diverged. This also seems unlikely because one end of the deletion in the human lineage would have to coincide precisely with the inversion point on the gibbon chromosome 16-homologous material. Because of the above reasons, it seems likely that the rearrangement occurred during a non-homologous end joining event, rather than an homologous recombination.

## 6.4 Discussion

For isolating the gibbon rearrangement point junctions, the vectorette PCR approach was rapid and successful, and has not been previously reported for use in cloning evolutionary rearrangement breakpoints. However, the sequence reads generated from the PCR products were short and, although it was possible to identify the homologous block junction points, it was impossible to interpret the genomic environment around the rearrangement event. In particular, the insertion of approximately 5 kb of DNA between the chromosome 22 and chromosome 16 homologous regions in one of the clones would have precluded the region's clear description using the vectorette approach. Therefore, a full sequence analysis was carried out on the breakpoint clones.

From the full sequence analysis of the gibbon rearrangement point clones and the HSA22 and HSA16 sequences homologous to the regions, which were broken in the gibbon lineage, it is not at this time possible to identify a specific rearrangement mechanism, although the evidence points to a non-homologous end-joining (NHEJ) event. During double strand break (DSB) repair via homologous recombination, the broken DNA sequence interacts with a homologous donor sequence and genetic information is exchanged between identical or nearly identical DNA sequences. NHEJ is accomplished by the joining of DNA ends without interaction between the broken molecule and a donor sequence. In NHEJ there is no requirement for homology at the DNA termini being joined, although NHEJ may be facilitated by short terminal homologies (Lin and Waldman, 2001). NHEJ may be accompanied by the deletion or gain of genetic material (for example, retrotransposon sequences) prior to healing of the DSB.

To take this work further, a detailed analysis of the relative orientation and degree of sequence similarity of the Alu elements flanking the rearrangement junction in 4_21 could be carried out by screening the sequences in RepeatMasker. This might indicate whether, due to the orientation of the elements, there was any possibility that these regions could have been involved in stabilising a homologous recombination event.