BIOINFORMATICS

Vol. 19 Suppl. 2 2003, pages ii162–ii171
DOI: 10.1093/bioinformatics/btg1074

Reconstructing tumor genome architectures

Benjamin J. Raphael[1,], Stanislav Volik[2], Colin Collins[2] and Pavel A. Pevzner[1]*

[1]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093-0114, USA and [2]Cancer Research Institute, University of California Comprehensive Cancer Center, San Francisco, CA 94115, USA

Received on March 17, 2003; accepted on June 9, 2003

ABSTRACT

Although cancer progression is often associated with genome rearrangements, little is known about the detailed genomic architecture of tumor genomes. The attempt to reconstruct the genomic organization of a tumor genome recently resulted in the development of the *End Sequence Profiling (ESP)* technique, and the application of this technique to human MCF7 tumor cells. We formulate the *ESP Genome Reconstruction Problem*, and develop an algorithm to solve this problem in the case of sparse ESP data. We apply our algorithm to analyze human MCF7 tumor cells, and obtain the first reconstruction of the putative architecture of human MCF7 tumor genome. Our results assist in the ongoing ESP analysis of MCF7 tumors by suggesting additional ESP experiments for the completion of a reliable reconstruction of the MCF7 tumor genome, and by focusing BAC re-sequencing efforts.

Contact: braphael@ucsd.edu

INTRODUCTION

Human cancer cells frequently possess large-scale chromosomal rearrangements. For example, in chronic myeloid leukemia, a chromosome 9;22 translocation causes the *ABL* gene on chromosome 9 to be brought under the regulation of the promoter of the *BCR* gene on chromosome 22 (Heisterkamp *et al.*, 1983). Many such rearrangements have been catalogued (Mitelman *et al.*, 2003), and the changes in gene structure and regulation caused by *individual* rearrangements have been intensively studied (Rowley, 1998). Recently, Sankoff *et al.* (2002) undertook the first computational analysis of individual rearrangements implicated in cancer.

Solid tumors are often associated with *multiple* (rather than single) chromosomal rearrangements. However, little is known about the detailed genomic architecture of tumor genomes, and the rearrangement scenarios associated with tumor genomes beyond information about some *single* reversals (also called inversions), translocations, fissions, and fusions. A variety of experimental techniques have been developed for the study of chromosomal rearrangements, including fluorescent *in situ* hybridization (Pinkel *et al.*, 1988; Thompson and Gray, 1993), chromosome painting (Jauch *et al.*, 1992), spectral karyotyping (Schröck *et al.*, 1996) and comparative genome hybridization (Kallioniemi *et al.*, 1992). However, the low resolution of these techniques is often a bottleneck in deriving the accurate architecture of a tumor genome. Genome-scale resequencing of a tumor genome would provide the ultimate dataset for cancer mutation and rearrangement studies; for example, a recent sequencing of a 3.2 MB region from 12 sporadic colon cancer cell lines revealed high mutation rates in this area (Wang *et al.*, 2002). However, this region is too short for comprehensive rearrangement studies, and it is currently infeasible to sequence an entire tumor genome in view of the high cost of mammalian genome sequencing. It is believed that cancer is manifested in both increased single nucleotide mutation rates and increased rate of genome rearrangement (Loeb *et al.*, 2003). Therefore, alternative technologies are sought for low-cost but high resolution mapping of tumor genome architectures and the associated analysis of genome rearrangements.

End Sequence Profiling (ESP) of a tumor genome is a recently developed experimental technique for rapid assessment of rearrangements in tumor cells (Volik *et al.*, 2003). ESP provides a balance between imprecise, but inexpensive technologies such as chromosome painting, and accurate, but expensive, genome sequencing. The ESP technique involves the construction of a BAC library for the tumor genome (the rearranged genome), the sequencing of the ends of BACs, and the mapping of end sequences to the human genome (the reference sequence). The technique has recently been applied to the first comprehensive rearrangement study of human MCF7 tumor cells (Volik *et al.*, 2003).

ESP follows the standard laboratory protocols for construction of a BAC library and the sequencing of BAC ends. The BAC end sequences (BES) are then mapped to

*To whom correspondence should be addressed.

the reference sequence using an algorithm such as BLAST (Altschul *et al.*, 1990). Typically, a BAC with insert size L will correspond to a pair of *BAC end sequences (BES pair)* separated by distance L in the reference genome. However, a BAC located in a region containing rearrangement breakpoints will result in a BES pair with locations in different parts of the reference genome—possibly on different chromosomes. We refer to such BACs as *composite* BACs. Each composite BAC suggests a rearrangement breakpoint internal to the composite BAC in the rearranged genome. For example, a single BAC whose ends map to different chromosomes in the reference genome suggests that a translocation occurred between these chromosomes in the rearranged genome. In this paper, we focus on the use of the mapped BES pairs to reconstruct the architecture of an entire tumor genome.

The analysis of rearrangements based on locations of BES pairs is complicated by the presence of *chimeric* BACs in the BAC library. Chimeric BACs are produced by joining of two non-contiguous regions of DNA, and ESP data are expected to have a significant number of chimeric BACs. Similar to composite BACs (containing the rearrangement breakpoints), the end sequences of chimeric BACs also map to widely separated regions of the reference genome. However, chimeric BACs are artifacts, rather than signs of real rearrangements. Therefore, chimeric BACs must be distinguished from composite BACs in order to derive the correct architecture of the rearranged genome. (See Fig. 1.)

In this paper, we formulate the problem of tumor genome reconstruction from ESP data and propose a heuristic algorithm to solve it in the case of sparse ESP data. We then apply the algorithm to ESP data from the human MCF7 tumor genome. We derive a putative genomic architecture of MCF7, and find 22 putative rearrangements associated with tumor progression. Three of the resulting rearrangement breakpoints are supported by existing experimental evidence, and additional experiments are currently underway to verify the remaining rearrangement breakpoints. Our reconstruction also suggests additional sequencing efforts (currently underway at the UCSF Cancer Center) that are required to complete a reliable reconstruction of the tumor genome. While a recent study (Volik *et al.*, 2003) derived several individual rearrangement breakpoints from ESP data, this paper presents the first attempt to derive the *entire* tumor genome architecture from ESP data.

ESP SORTING PROBLEM

In this section, we formalize the problem of reconstruction of the architecture of the rearranged genome based on the ESP data. An ESP experiment assumes that the reference sequence, G is known, and that a BAC library for the (unknown) rearranged genome G' is constructed. G' is assumed to evolve from G by a series of reversals, translocations, fusions, and fissions. The ESP experimental data provides the locations in G of the BES pairs for a set of BACs from the BAC library for G'. The problem is to derive the genomic architecture of the rearranged genome, G', from this data.

First, we shall formulate a simplified version of the ESP genome reconstruction problem for a unichromosomal genome G—represented by the interval $[0, 1]$—under the assumption that there are no chimeric BACs in the BAC library. Since the length of BAC end sequences—usually 500–700 nucleotides—is small at the genome-scale, we represent a BES as a point in G. A BAC corresponds to a BES pair $(x, y) \in G \times G$, where we order the BES's in the pair such that $x < y$. We fix a constant L, which is the maximum insert length of a BAC in the library. We say that a BES pair (x, y) is *valid* if $y - x \leq L$, and *invalid* otherwise. BES pairs of most BACs form valid pairs while BES pairs of composite BACs typically form invalid pairs (Fig. 1).

For each $s, t \in [0, 1]$ with $s < t$, we define the *reversal* $\rho_{s,t} : [0, 1] \mapsto [0, 1]$ by

$$\rho_{s,t}(x) = \begin{cases} x, \text{ if } x < s, \text{ or } x > t, \\ t - (x - s), \text{ otherwise.} \end{cases}$$

A sequence of genome rearrangements that transforms G into G' also transforms the BES positions correspondingly. In particular, the invalid pair (x, y) in $G \times G$ may be transformed into the valid pair $(\rho x, \rho y)$ in the rearranged genome $G' = \rho G$, where $\rho = \rho_{s_1,t_1} \cdots \rho_{s_k,t_k}$ denotes a series of reversals. Since all BES pairs in the rearranged genome G' are valid (under our assumption that there are no chimeric BACs in the BAC library), the rearrangement scenario $G \mapsto G'$ can be viewed as the process of elimination of invalid BES pairs. Figure 2 shows an example of the simplified ESP data, and two reversals that make all BES pairs valid. We remark that for a given set of BES pairs, a rearranged genome with all BES pairs valid is not necessarily unique (Fig. 2b).

With this framework, we consider the following problem.

ESP SORTING PROBLEM. *Given BES pairs* (x_1, y_1), $\ldots$, (x_n, y_n), *find the minimum number of reversals* $\rho_{s_1,t_1}, \ldots, \rho_{s_k,t_k}$ *such that if* $\rho = \rho_{s_1,t_1} \cdots \rho_{s_k,t_k}$ *then* $(\rho x_1, \rho y_1), \ldots, (\rho x_n, \rho y_n)$ *are valid pairs.*

One can always sort n BES pairs with at most n reversals by bringing elements (x_i, y_i) within distance smaller than L in every step. In order to solve the ESP Sorting Problem with fewer than n reversals, one has to use some reversals that eliminate more than one

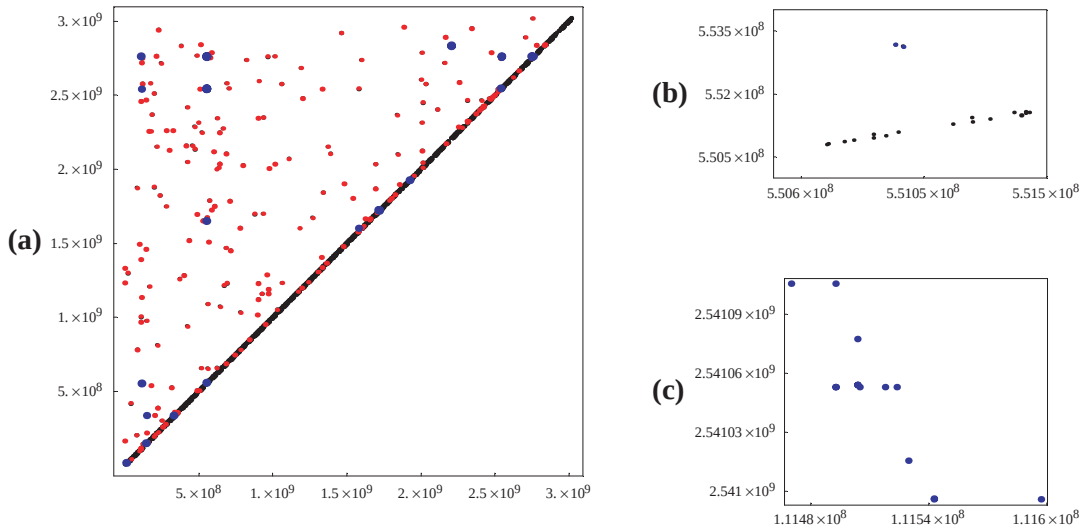

Fig. 1. (a) ESP data from the MCF7 tumor genome (June 1, 2003 dataset). Each point (x, y) corresponds to a BES pair, where x and y are the genomic coordinates in the human genome of the first and second read from the pair. The chromosomes are concatenated to form a single coordinate system, and for illustrative purposes, points are drawn in exaggerated scale. With maximum insert length $L = 200$ kb, 5856 out of a total of 6239 BES pairs satisfy the BAC length constraint (black points), while the remaining 383 invalid BES pairs correspond to composite and chimeric BACs. 256 out of 383 invalid BES pairs are isolated (red points), while 127 of the remaining invalid BES pairs form 30 clusters that suggest composite BACs (shown by slightly enlarged blue points that represents two or more invalid BES pairs). 5 out of 30 clusters (containing 15 BES pairs) are located near the main diagonal with $y - x$ varying from $L = 200$ kb to 1.2 Mb. Such BES pairs may be signs of microrearrangements in the tumor genome (compare with Pevzner and Tesler (2003a)). Note that due to scaling issues, multiple clusters may appear as a single blue point in the figure. (b) Expanded view of region from chromosome 3. The two blue points form a cluster that appeared as a single blue point in (a). (c) Expanded view of a cluster of BES pairs that indicate a chromosome 1;17 translocation.

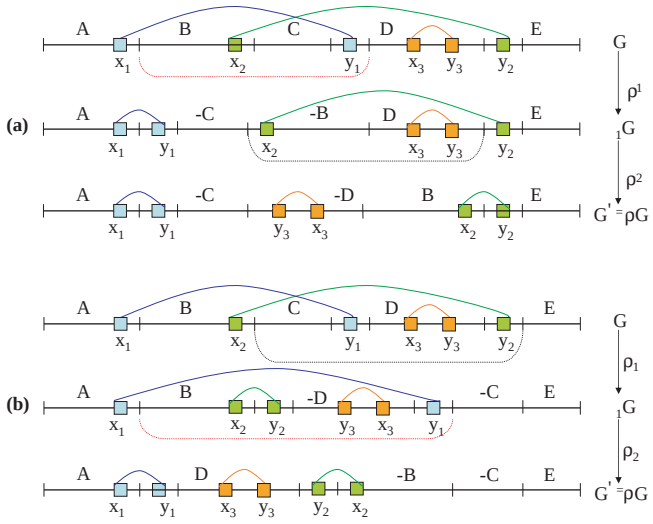

Fig. 2. Schematic view of ESP data with three BES pairs shown as arcs connecting BES elements represented as colored squares. (a) A sequence of two reversals ρ_1, ρ_2 produces $G' = \rho G$ with all BES pairs valid, where $\rho = \rho_1 \cdot \rho_2$. (b) A different sequence of two reversals $\tilde{\rho_1}, \tilde{\rho_2}$ produces $G' = \tilde{\rho} G$ with all BES pairs valid, where $\tilde{\rho} = \tilde{\rho_1} \cdot \tilde{\rho_2}$. Thus, more than one rearrangement scenario is consistent with ESP data.

invalid BES pair in a single step. We say that BES pairs (x_1, y_1) and (x_2, y_2) are *correlated* if $|x_2 - x_1| + |y_2 - y_1| \leq 2L$. It is easy to see that if (x_1, y_1) and (x_2, y_2) are invalid BES pairs and there exists $\rho = \rho_{s,t}$ such that $(\rho x_1, \rho y_1)$ and $(\rho x_2, \rho y_2)$ are valid, then (x_1, y_1) and (x_2, y_2) and correlated. The converse is also true is certain cases, and one can construct examples of chains of correlated pairs. Our heuristic approach to analyzing MCF7 data is based on the observation that composite BACs sharing rearrangement breakpoints for the same reversals/translocations typically produce correlated BES pairs.[†] If an ESP project has a high BAC coverage, most pairs of rearrangement breakpoints are covered by two or more BACs, providing the possibility to recover pairs of breakpoints that describe all (or almost all) individual rearrangements. However, information about pairs of 'correlated' breakpoints is not sufficient for reconstruction of genomic architecture. For example, a correlated pair

[†] At the time of writing this paper, the complexity status of the ESP Sorting Problem remains unknown. However, the sparse nature of the existing ESP data allowed us to come up with a heuristic that works well and leads to provably optimal solutions for existing ESP data. At the same time, our heuristic approach might fail in the case of ESP experiments with tens/hundreds of thousands of BACs, and more rearranged genomes than MCF7.

of breakpoints residing on different chromosomes could arise from two possible translocations. Below, we describe how the *oriented* ESP data help to resolve the genome reconstruction problem.

ESP GENOME RECONSTRUCTION PROBLEM

Our focus in this paper is the analysis of the MCF7 tumor genome, and consequently we extend the above simple formulation of the ESP Sorting Problem to take into account the following features of real ESP datasets:

1. G is a multichromosomal, rather than unichromosomal genome.
2. In addition to reversals, the set of genome rearrangements includes (frequent) translocations, and (less frequent) fissions and fusions.
3. Each BES possesses an orientation, which corresponds to the strand (plus or minus) where the BES is mapped. BAC end sequencing produces 'convergent' BES pairs in G' with opposite orientations, i.e. elements of a BES pair come from opposite strands and are directed towards each other.

To reflect the orientation of BES's, we assign plus and minus signs to each element of a BES pair. When a BES is matched to the reference genome, the match may occur either on the positive or negative strand. A BES whose match is on the positive (resp. negative) strand is given a positive (resp. negative) orientation. The orientations of BES has important consequences for the genome reconstruction problem, since they eliminate the ambiguities in deriving translocations from the set of correlated breakpoints and allow a reliable reconstruction of the architecture of the rearranged genome. Although single rearrangements in tumor genomes were analyzed in depth by different groups in the past, the possibility of the reconstruction of the entire tumor genome escaped the attention of previous researchers.

We represent a BES position as either $+x$ or $-x$, depending on its orientation. The typical non-composite BAC has BES pair $(+x, -y)$, where $|y| \approx |x| + L$, reflecting the fact that the corresponding read pairs have opposite and converging directions. Every BES pair (x, y) in ESP data possesses one of four possible orientations: $(+,+)$, $(+,-)$, $(-,+)$, $(+,+)$.

We concatenate chromosomes in the multichromosomal genome G into a virtual unichromosomal genome—represented by the interval $[0, 1]$, and we define the distance between positions x and y in the resulting concatenate by

$$d(x,y) = \begin{cases} |\,|y| - |x|\,|, \\ \quad \text{if } x \text{ and } y \text{ lie on the same chromosome,} \\ \infty, \text{ otherwise.} \end{cases}$$

The distance between elements x and y of BES pair (x, y) is now defined as $d(x, y)$.

For positions s and t in G with $s < t$, a reversal $\rho_{s,t}$ acts on a (signed) position x as follows:

$$\rho_{s,t}(x) = \begin{cases} x, \text{ if } x < s \text{ or } x > t \\ -\operatorname{sign}(x)\,(t - (|x| - s))\,, \text{ otherwise.} \end{cases}$$

This transformation is similar to the transformation described in the previous section for unsigned x, with the only difference being that we now take into account that reversals also flip the signs of oriented positions. This operation describes reversals $\rho_{s,t}$ and does not cover translocations, fusions, and fissions. However, translocations, fissions, and fusions can be modelled as reversals in the concatenated genome (Hannenhalli and Pevzner, 1995). This reduction of a multichromosomal genome to a unichromosomal genome has to be done carefully since it restricts some types of rearrangements. To address this issue we allow an entire chromosome to be flipped 'for free' in our approach to the ESP Genome Reconstruction Problem. Since the number of putative rearrangements in our scenario is relatively small, we can afford flipping and rearranging the chromosomes in a concatenate when necessary. For the sake of simplicity, below we assume that the concatenate is fixed.

Since the experimental protocol of end sequencing uses primers from opposite DNA strands, our definition of valid BES pairs now must include the requirement that the orientations of the elements in the pair are opposite and convergent. An oriented BES pair (x, y) is *valid* if $|y| - |x| \leq L$, $x > 0$ and $y < 0$, (i.e. $\operatorname{sign}(x) = '+'$, and $\operatorname{sign}(y) = '-'$.) Using this framework, we consider the following problem.

ESP GENOME RECONSTRUCTION PROBLEM. *Given a set of oriented BES pairs generated by an ESP experiment, identify composite BES pairs* $(x_1, y_1), \ldots, (x_n, y_n)$ *in this set, and find the minimum number of rearrangements* $\rho_{s_1,t_1}, \ldots, \rho_{s_k,t_k}$ *such that if* $\rho = \rho_{s_1,t_1} \cdots \rho_{s_k,t_k}$ *then* $(\rho x_1, \rho y_1), \ldots, (\rho x_n, \rho y_n)$ *are valid pairs.*

We illustrate the importance of the orientations of the BES pairs for genome reconstruction with an example. Consider two correlated invalid BES pairs whose ends lie on different chromosomes. Such a configuration suggests a translocation ρ with breakpoints near the ends of the BES pairs (Fig. 3a). However, the locations of the breakpoints of such a translocation relative to the BES pairs is not known. Furthermore, the locations of the breakpoints of this translocation are not sufficient to determine the translocation uniquely. If A, B and C, D represent two chromosomes with segments labelled by A, B, C, and D, then two translocations with the same breakpoints AB

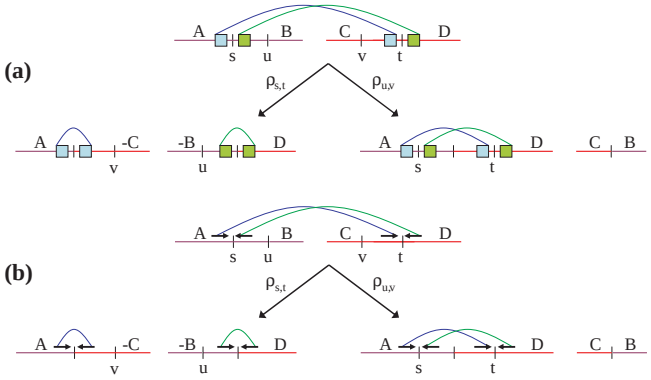

Fig. 3. ESP data on two chromosomes reveals the importance of BES orientations. (a) With unsigned BES pairs, there exist two translocations that transform invalid BES pairs into valid pairs. Unsigned ESP data do not allow one to identify which translocation is real. (b) The signs of BES pairs (indicated by arrows) determine the translocation that transforms the invalid BES pairs into valid pairs. The translocation on the right does not produce valid BES pairs, and is classified as an artifact.

and CD are possible: one translocation results in chromosomes A, D and C, B; the other translocation results in chromosomes $A, -C$ and $-B, D$. Because of these ambiguities we cannot infer the order of segments in the rearranged genome solely from the *locations* of the BES pairs. The simple, but important insight is that the *orientations* of the BES pairs in the cluster determine the appropriate translocation (Fig. 3b). Hence, to derive a plausible reconstruction of the architecture of the rearranged genome, it is essential to consider the orientations of the BES pairs.

IDENTIFYING COMPOSITE BES PAIRS

The technique that we use to separate chimeric and composite BACs is similar in spirit to the technique used in DNA sequencing to detect chimeric reads (Pevzner *et al.*, 2001). Chimeric reads plague every DNA sequencing project and all existing fragment assembly tools detect and remove chimeric reads. These algorithms are based on the assumption that chimeric reads typically combine two 'random' segments of DNA and therefore different chimeric reads rarely use DNA from closely located genomic regions. Similarly, distinct chimeric BACs rarely use DNA material from closely located genomic regions. On the other hand, composite BACs containing the same rearrangement breakpoint often come in clusters that use DNA from closely located genomic regions.

Figure 1 illustrates that while most BES pairs (x, y) lie in a band of width L around the diagonal (valid BES pairs), there are also many invalid BES pairs arising from either chimeric or composite BACs.[‡] Many of the invalid BES pairs (red points in Fig. 1) are isolated; i.e. there is no other BES pair within distance $2L$ from them. However, a small number of BES pairs form clusters (blue points in Fig. 1). Every blue point corresponds to a cluster of correlated BES pairs, as shown in the expanded view (Fig. 1b and c).

We define the distance between BES pairs (x_1, y_1) and (x_2, y_2) as $d(x_1, x_2) + d(y_1, y_2)$. Given a set of BES pairs, we say that BES pair (x, y) is *isolated* if its distance from all other BES pairs in the set is larger than $2L$. With the ESP data alone, it is difficult to determine whether isolated invalid pairs arise from composite BACs or chimeric BACs, particularly in the case of low BAC coverage. We will assume that isolated BES pairs arise from chimeric BACs. With this assumption, we may incorrectly classify some composite BACs as chimeric, but such misclassification errors decrease as the BAC coverage increases. However, our chance of classifying a chimeric BAC as composite is extremely small. Hence, we identify composite BACs by removing all isolated invalid BES pairs (chimeric BACs), and combining the remaining invalid BES pairs into clusters, where the BES pairs in each cluster are pairwise correlated (putative composite BACs).

ESP GENOME RECONSTRUCTION APPROACH

Following filtering of isolated invalid BES pairs, we are left with K clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$ of invalid BES pairs. The problem now is to find the minimum number of rearrangements $\rho_1, \ldots, \rho_k$ such that if $\rho = \rho_1 \cdots \rho_k$ then $(\rho x, \rho y)$ is a valid BES pair for all $(x, y) \in \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_K$. In this section, we present a heuristic approach to this problem that works well with existing *sparse* ESP datasets. This approach uses the locations and orientations of the BES pairs in the clusters to both define a set of 'synteny blocks' in the reference genome G, and determine the order of these blocks in the rearranged genome G'.

One might consider the following naive approach to the ESP Genome Reconstruction Problem: for each cluster $\mathcal{C}_i$, find a rearrangement ρ_i that transforms all invalid pairs in $\mathcal{C}_i$ into valid pairs. Such a rearrangement does not always exist. However, even if such a rearrangement exists, the naive approach is unlikely to lead to a correct reconstruction of G', except in the special case of 'non-overlapping' rearrangements. The reason that the naive approach fails is because the architecture of G' depends on the *order* in which the rearrangements are performed. Consequently, the genomic architecture obtained by this naive approach will depend on the order in which we de-

[‡] Since L is small on the genome scale, valid BES pairs in Figure 1 appear to be positioned exactly on the diagonal, rather than within a band of width $2L$.

fine rearrangements from the clusters. The problem, therefore, is to determine the architecture of G' solely from the clusters *without* knowing the order of the rearrangements. We now describe a heuristic for this problem that works well under certain reasonable assumptions.

We assume that the ESP data is *sparse*, i.e. no BAC contains more than one rearrangement breakpoint and each cluster results from a single rearrangement. We cannot determine this rearrangement directly from the cluster, since we do not know the order of rearrangements that produced G'. However, we can use the requirement that in G' all BES pairs in the cluster are valid to define the boundaries of 'synteny blocks' in G and determine how these blocks are connected in G'. Specifically, for a cluster $\mathcal{C} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ consisting of m invalid BES pairs, consider a pair of breakpoints s and t in G. The points s and t divide G into blocks $A = [0, s)$, $B = [s, t]$, and $C = (t, 1]$ (Fig. 4a). The goal is to define s and t and to determine how the ends of the blocks A, B, and C should be connected in G' in order to transform the invalid BES pairs $(x_1, y_1), \ldots, (x_m, y_m)$ into valid pairs in G'. Additionally, we require that any valid pairs (x, y) overlapping an element of $\mathcal{C}$ (e.g. $x < x_i < y$ for some i) remain valid in G'.

We first determine approximate locations for s and t by recalling that in order for the BES pairs in $\mathcal{C}$ to be valid in G', they must have opposite, convergent orientations. Consequently, each x_i must be oriented toward s and each y_i must be oriented toward t. This requirement determines an interval S, containing s, and an interval T containing t. Additionally, the requirement of opposite, convergent orientations also determines how the ends of the blocks A, B, and C are connected in G' (Fig. 4a). After determining intervals S and T, the requirement that valid BES pairs (x, y) must satisfy the BAC length constraint $|y - x| \leq L$ implies that s and t must satisfy $\max_{(x,y)\in\mathcal{C}} |x - s| + |y - t| \leq L$ (Fig. 4b). We compute the rearrangement breakpoints $s \in S$ and $t \in T$ that minimize the maximum length of BES pairs in G'; i.e. we solve

$$\min_{s\in S, t\in T} \max_{(x,y)\in\mathcal{C}} |x - s| + |y - t|.$$

Thus, from a cluster $\mathcal{C}$, we define the boundaries of blocks A, B, and C in G, and determine how the ends of these blocks are connected in G'.

The K pair of breakpoints defined from the clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$ divide G into $2K + 1$ synteny blocks, and the clusters determine how the ends of these blocks are connected in G'. The final step of the genome reconstruction is to combine this information—derived from the individual clusters—into the architecture of G'. Figure 5 shows an example with three unknown reversals transforming the 'human genome' A, B, C, D, E, F, G into a 'tumor genome' $A, -C, F, -D, B, -E, G$, with the seven synteny blocks defined by three clusters $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$. In real ESP experiments, the three reversals and the architecture of the tumor genome are unknown and the goal is to derive the architecture from the ESP data, i.e. from the clusters $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$. Given synteny blocks U and V in G, we say that (U, V) is an *oriented breakpoint* in G' if the end of block U is connected to the start of block V in G'. Clusters $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$ reveal three pairs of oriented breakpoints in the tumor genome (Fig. 5b). Superimposing the three oriented breakpoint pairs (Fig. 5c) reveals the synteny block order in the tumor genome as a simple path through the synteny blocks (Fig. 5d).

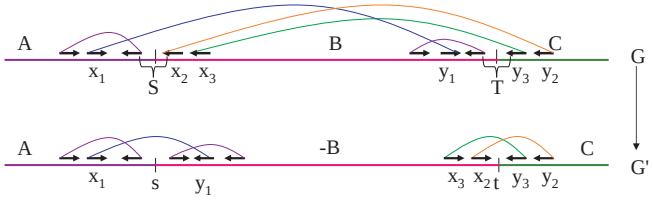

Fig. 4. ESP Genome Reconstruction: resolving a cluster $\mathcal{C} = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$ of invalid BES pairs. Top: The cluster $\mathcal{C}$ in G. In G', the elements of each pair (x_i, y_i) must have opposite, convergent orientations. This requirement, and consideration of overlapping valid BES pairs (purple arcs) determine an interval S containing breakpoint s, and an interval T containing breakpoint t. Breakpoints s and t divide G into blocks A, B, C. Bottom: The orientation of BES pairs in $\mathcal{C}$ determine that the end of block A is connected to the end of block B in G'. Similarly, the start of block B is connected to the start of block C in G'. Breakpoints s and t are defined such that $\max_{(x,y)\in\mathcal{C}} |x - s| + |y - t| \leq L$.

We say that the ESP data is *complete* if each rearrangement breakpoint in the rearranged genome is internal to a BAC from the BAC library, i.e. there are no 'silent' rearrangements that do not disrupt any BACs. This assumption holds whenever the coverage of the end sequenced BACs is sufficiently large. When the ESP data is *complete*, the graph obtained from the superposition of the oriented breakpoint pairs corresponds to the breakpoint graph (Pevzner, 2000) for the signed permutation of the synteny blocks in the rearranged genome (Fig. 5e). To transform the graph in Figure 5c into the breakpoint graph, one has to remove edges corresponding to the synteny blocks and add edges corresponding to the breakpoint regions.

The ESP genome reconstruction approach just described requires, for each cluster $\mathcal{C}$, the determination of a pair of breakpoints and a connection between the ends of the synteny blocks adjacent to these breakpoints in G' such that *all* the BES pairs in the cluster are valid in the G'. For the ESP data from the MCF7 tumor genome, this is possible for 22 of the 25 clusters. In the next section, we describe the complications that arise for the remaining 3 clusters.

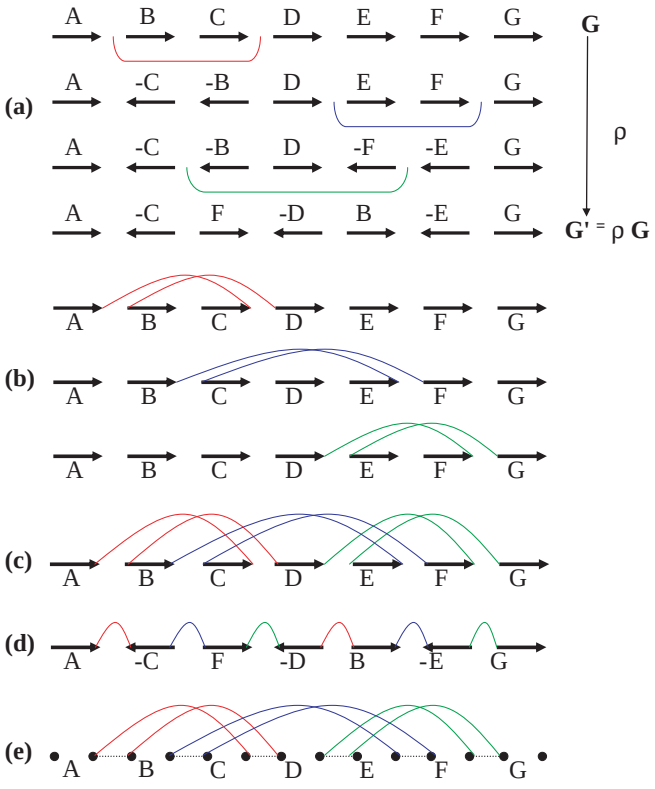

Fig. 5. (a) Transformation of 'human' genome A, B, C, D, E, F, G into a 'tumor' genome $A, -C, F, -D, B, -E, G$ by three reversals. (b) Clusters in the ESP data will reconstruct six breakpoints $(A, -C), (-C, F), (F, -D), (-D, B), (B, -E), (-E, G)$ arranged into three pairs: $(A, -C) \oplus (-D, B)$, $(-C, F) \oplus (B, -E)$ and $(F, -D) \oplus (-E, G)$ corresponding to the three reversals. (b) Top: modelling of oriented breakpoint pair $(A, -C) \oplus (-D, B)$. Middle: modelling of oriented breakpoint pair $(-C, F) \oplus (B, -E)$. Bottom: modelling of oriented breakpoint pair $(F, -D) \oplus (-E, G)$. (c) Superimposing all edges corresponding to oriented breakpoint pairs. The resulting graph is a simple path through the synteny blocks that provides the reconstruction (d) of the tumor genome. (e) The breakpoint graph for the signed permutation of synteny blocks in the tumor genome. (Note that this graph differs from the canonical construction of a breakpoint graph.)

RESULTS

We apply the genome reconstruction approach from the previous section to ESP data from MCF7 tumor cells generated at the University of California, San Francisco Cancer Center. The BAC library for MCF7 consists of approximately 68 000 BACs and large-scale ESP efforts are currently underway to analyze this library. While Volik *et al.* (2003) were able to derive some rearrangements from ESP data, the number of rearrangements and the genomic architecture of the tumor genome remain unknown. The ESP data (as of June 1, 2003) includes 6239 BACs with uniquely mapped BAC end sequences of roughly 500 bp. New BES pairs continue to be sequenced and mapped, and will be incorporated into the genomic reconstruction. We use the NCBI April 2003 build of the human genome as the reference genome.

Setting $L = 200$ kb, we find 383 invalid BES pairs, 127 of which form 30 clusters. If these 383 invalid BES pairs were placed randomly in the human genome, the probability of a cluster of at least two BES pairs is approximately 1.3×10^{-3} (Glaz *et al.*, 2001). We conclude that the 127 BES pairs in these 30 clusters are likely to represent composite BACs, and use these clusters to find the architecture of the tumor genome. However, for 15 of these 127 BES pairs (corresponding to 5 clusters), the distance between the elements of the BES pair is less than 1.2 Mb, suggesting that they correspond to potential microrearrangements that have only minor influence on the large-scale genomic architecture of the tumor genome. Microrearrangements are not widely reported in the literature on chromosomal aberrations, probably because cytogenetic techniques do not allow reliable identification of microrearrangements. However, it is premature to claim that microrearrangements are frequent in tumor cells from this data, since some of these microrearrangements may be caused by assembly errors.

We applied our ESP genome reconstruction approach to the remaining 25 clusters, which contain 112 invalid BES pairs. For 3 of the 25 clusters, we were unable to find a rearrangement that transforms *all* BES pairs in the cluster into valid pairs (see below). Table 1 (**supplementary information**) lists the derived locations of the breakpoints for the remaining 22 clusters. The cytogenetic locations of 3 breakpoint pairs correspond to previously observed chromosomal aberrations in tumors using FISH and SKY (Volik *et al.*, 2003). Many of the breakpoints lie in regions 3p14, 17q23, and 20q12–13, and there is experimental evidence that these regions are duplicated several times in the MCF7 genome (Bärlund *et al.*, 2002; Volik *et al.*, 2003).

We suspect that the difficulty with the 3 clusters that we failed to resolve arises from extensive duplications in the tumor genome. Indeed, all of these unresolved clusters contain BES pairs from the duplicated region 20q12–20q13. This region is particularly prominent with the current ESP dataset because the end-sequenced BACs were enriched for BACs from this region, using specific BAC selection protocols (Volik *et al.*, 2003). In duplicated regions, our assumption of sparse data may be violated, and our algorithm requires modifications to handle this case. We expect that multiple pairs of breakpoints from different instances of the duplicated region in the tumor genome will map to the same cluster of BES pairs from this region. Unfortunately, the low coverage of the current ESP data does not permit us to resolve multiple breakpoints in the duplicated regions.

Table 1. Breakpoints in the MCF7 tumor genome derived from the current ESP data

Breakpoint 1	Breakpoint 2	Cytogenetic Location
(1, 107083885)	(20, 53176720)	**1p13.3;20q13.2**
(1, 111469938)	(17, 57403767)	1p13.2;17q23.2
(1, 114244003)	(3, 62893420)	1p13.2;3p14.2
(1, 142545502)	(1, 145692028)	1q21.1;1q21.2
(1, 145899288)	(2, 91285936)	1q21.2;2p11.2
(3, 61695632)	(9, 113077983)	3p14.2;9q33.1
(3, 62454840)	(3, 64601224)	3p14.2;3p14.1
(3, 63593257)	(17, 59179130)	3p14.2;17q23.2
(3, 63821304)	(17, 57029534)	3p14.1;17q23.2
(3, 63922405)	(20, 47112585)	3p14.1;20q13.13
(3, 64793126)	(20, 54193973)	**3p14.1;20q13.2**
(9, 42803353)	(9, 61133368)	9p11.2;9q12
(10, 46058619)	(10, 51147527)	10q11.21;10q11.23
(10, 47835601)	(10, 51699199)	10q11.22;10q11.23
(14, 18138658)	(22, 14764022)	14q11.2;22q11.1
(17, 59651864)	(20, 50068722)	17q23.2;20q13.13
(17, 59671150)	(20, 54673221)	**17q23.2;20q13.2**
(17, 60300954)	(20, 56318032)	17q23.2;20q13.31
(20, 41456347)	(20, 52907427)	20q12;20q13.2
(20, 41649176)	(20, 53366067)	20q12;20q13.2
(20, 46409526)	(20, 56429084)	20q13.12;20q13.31
(20, 52839025)	(20, 56477622)	20q13.2;20q13.31

Columns list the chromosomal coordinates of each breakpoint and the corresponding cytogenetic location. Breakpoints with supporting experimental evidence are indicated in bold type.

From our genome reconstruction approach, we obtain a putative reconstruction of the MCF7 tumor genome (Fig. 6). We emphasize that it is a preliminary reconstruction that does not include microrearrangements and is likely to miss some rearrangements: either due to low coverage of existing ESP data, or due to complications with deletions/duplications. In fact, the reconstruction of the MCF7 genome that we obtain excludes 5 synteny blocks of total length 27 Mb, suggesting that the ESP data is not complete. While deletions can explain the exclusion of these blocks, more likely 'silent' rearrangements with breakpoints inside these blocks are the cause. The locations of these blocks in the human genome suggest which additional ESP experiments will be necessary to incorporate these blocks into the MCF7 genome reconstruction.

We remark that both the ESP Sorting Problem and the ESP Genome Reconstruction Problem become rather difficult in the presence of duplicated regions, and the ultimate explanation for the unresolved clusters in the MCF7 dataset will be provided by rather time-consuming BAC resequencing efforts.

FUTURE DIRECTIONS

Our analysis of the MCF7 genome yields the first high-resolution (albeit incomplete) picture of the genomic architecture of a complex tumor genome. We expect that this picture will further improve as the number of mapped BES pairs increases. However, even with the current dataset, we have strong evidence of 19 previously unknown rearrangements. The experiments to verify the corresponding breakpoint pairs are currently underway. See Raphael *et al.* (2003) for further details. In addition, some BACs that contain putative rearrangement breakpoints are being fully sequenced to determine the exact breakpoint locations. Our results will provide assistance in guiding the resequencing effort.

Duplications and deletions appear to be common events in the progression of tumor genomes. One mechanism of duplication is the enlargement of repetitive sequences (microsatellites) due to defects in mismatch repair genes in tumors (Fishel *et al.*, 1993; Ionov *et al.*, 1993). In the ESP dataset from the MCF7 tumor, there is strong evidence to suggest both duplications and deletions. We intend to include duplications and deletions in the framework of the ESP Genome Reconstruction Problem, to extend our algorithm to the case of non-sparse data, and to study further the complexity of the ESP Sorting Problem.

While our reconstruction of the MCF7 tumor genome is preliminary, a more complete reconstruction may shed light on certain questions about genome evolution, such as the genomic distribution of rearrangement breakpoints associated with cancer, and the correlation/independence of cancer breakpoints and 'evolutionary' breakpoints. Sankoff *et al.* (2002) demonstrate that for a number of clinically determined rearrangements in human cancers from the database of Mitelman *et al.* (2003), the breakpoints tend to be clustered in the medians of the chromosomal arms. However, they remark that it is unclear if this observation a genuine biological phenomenon or a result of bias in the data collection. Some researchers hypothesize that the breakpoints of cancer rearrangements are clustered at 'fragile' sites in the genome (Smith *et al.*, 1998; Miró *et al.*, 1987). Dunham *et al.* (2002) recently observed a similar clustering of breakpoints in genome rearrangements of *Saccharomyces cerevisiae* that were subjected to various selective pressures in the laboratory. A more complete reconstruction of the entire architecture of the MCF7 genome may contribute to the understanding of the distribution of cancer breakpoints.

Nadeau and Taylor (1984) proposed that the breakpoints of evolutionary rearrangements are randomly distributed across the human genome. The breakpoint distributions in comparative maps of human and mouse support this hypothesis (Sankoff *et al.*, 2002, 1997). On the other hand, comparisons of the recently available whole genome sequences of human and mouse suggest that breakpoints of evolutionary rearrangements are clustered (Pevzner and Tesler, 2003b). Once we obtain an accurate reconstruction of the MCF7 tumor genome, we will examine whether

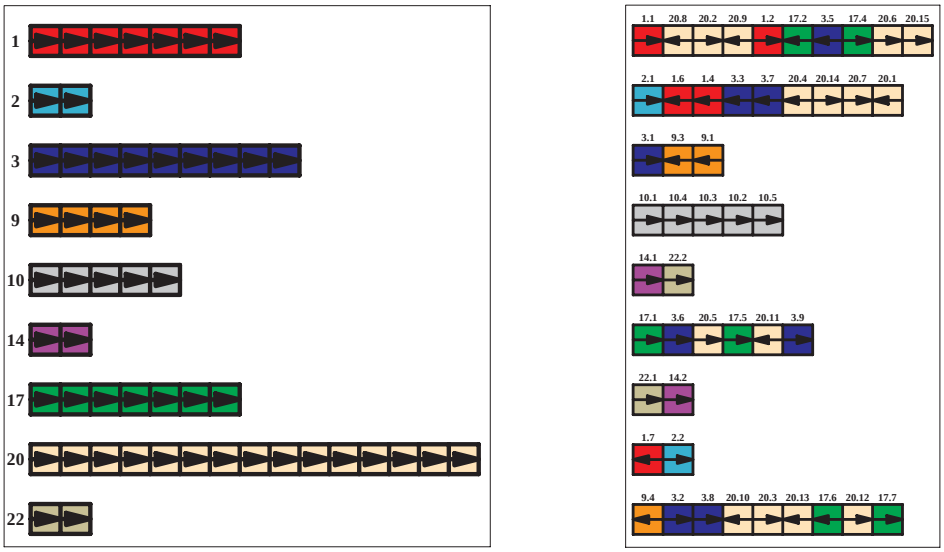

Fig. 6. The nine human chromosomes (left) that are rearranged in the MCF7 tumor genome (right), as derived from current ESP data. Other chromosomes are unchanged. Genomic blocks are color coded and oriented (indicated by arrows) according to their chromosomal locations and orientation in the human genome. Using the GRIMM program (Tesler, 2002), we find parsimonious rearrangement scenario that produces the MCF7 genome by a sequence of 5 reversals and 15 translocations.

the locations of MCF7 breakpoints are correlated with the breakpoints in the human-mouse genomes. As a first step, we examined the 22 breakpoints that we derived in the MCF7 genome with those in the human-mouse genome comparisons. Some of the tumor breakpoints fall inside or close to the human-mouse breakpoint regions. However, with the small number of MCF7 breakpoints, it remains to be seen whether this is a real or chance correlation, and further investigation is required.

ACKNOWLEDGEMENTS

We are grateful to Glenn Tesler for the analysis of correlations between evolutionary and tumor breakpoints, and to Haixu Tang for helpful discussions. B.R. was supported by an Alfred P. Sloan Postdoctoral Fellowship. The work of S.V. and C.C. was performed with support from the Department of Defense grant DAMD100110500, Breast Cancer Research Program grant 8WB-0054, Bay Area Breast Cancer Special Project Of Research Excellence CA58207 and the Avon Foundation. The work of B.R. and P.P. was supported by National Institutes of Health grant 1 R01 HG02366.

REFERENCES

Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bärlund,M., Monni,O., Weaver,J., Kauraniemi,P., Sauter,G., Heiskanen,M., Kallioniemi,O. and Kallioniemi,A. (2002) Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. *Genes Chromosomes Cancer*, **35**, 311–317.

Dunham,M., Badrane,H., Ferea,T., Adams,J., Brown,P.O., Rosenzweig,F. and Botstein,D. (2002) Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **99**, 16144–16149.

Fishel,R., Lescoe,M., Rao,M., Copeland,N., Jenkins,N., Garber,J., Kane,M. and Kolodner,R. (1993) The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell*, **75**, 1027–1038.

Glaz,J., Naus,J. and Wallenstein,S. (2001) *Scan Statistics*. Springer.

Hannenhalli,S. and Pevzner,P. (1995) Transforming men into mice (polynomial algorithm for genomic distance problem. In *Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science*. Milwaukee, Wisconsin, pp. 581–592.

Heisterkamp,N., Stephenson,J., Groffen,J., Hansen,P., de Klein,A., Bartram,C. and Grosveld,G. (1983) Localization of the c-ab1 oncogene adjacent to a translocation break point in chronic myelocytic leukaemia. *Nature*, **306**, 239–242.

Ionov,Y., Peinado,M., Malkhosyan,S., Shibata,D. and Perucho,M. (1993) Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature*, **363**, 558–561.

Jauch,A., Wienberg,J., Stanyon,R., Arnold,N., Tofanelli,S., Ishida,T. and Cremer,T. (1992) Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. *Proc. Natl Acad. Sci. USA*, **89**, 8611–8615.

Kallioniemi,A., Kallioniemi,O., Sudar,D., Rutovitz,D., Gray,J., Waldman,F. and Pinkel,D. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Loeb,L., Loeb,K. and Anderson,J. (2003) Multiple mutations and cancer. *Proc. Natl Acad. Sci. USA*, **100**, 776–781.

Miró,R., Clemente,I., Fuster,C. and Egozcue,J. (1987) Fragile sites, chromosome evolution, and human neoplasia. *Hum. Genet.*, **75**, 345–349.

Mitelman,F., Johansson,B. and Mertens,F. (eds.) (2003) *Database of Chromosome Aberrations in Cancer*. http://cgap.nci.nih.gov/Chromosomes/Mitelman.

Nadeau,J.H. and Taylor,B.A. (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl Acad. Sci. USA*, **81**, 814–818.

Pevzner,P. (2000) *Computational Molecular Biology: An Algorithmic Approach*. MIT Press.

Pevzner,P. and Tesler,G. (2003a) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res.*, **13**, 37–45.

Pevzner,P. and Tesler,G. (2003b) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl Acad. Sci. USA*, **100**, 7672–7677.

Pevzner,P.A., Tang,H. and Waterman,M.S. (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl Acad. Sci. USA*, **98**, 9748–9753.

Pinkel,D., Landegent,J., Collins,C., Fuscoe,J., Segraves,R., Lucas,J. and Gray,J. (1988) Fluorescence *in situ* hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proc. Natl Acad. Sci. USA*, **85**, 9138–9142.

Raphael,B., Volik,S., Colins,C. and Pevzner,P. (2003) Reconstruction of the genomic architecture of a breast cancer tumor. *(preprint)*.

Rowley,J. (1998) The critical role of chromosome translocations in human leukemias. *Annu. Rev. Genet.*, **32**, 495–519.

Sankoff,D., Deneault,M., Turbis,P. and Allen,C. (2002) Chromosomal distributions of breakpoints in cancer, infertility, and evolution. *Theor. Popul. Biol.*, **61**, 497–501.

Sankoff,D., Parent,M.N., Marchand,I. and Ferretti,V. (1997) On the Nadeau–Taylor theory of conserved chromosome segments. In Apostolico,A. and Hein,J. (eds), *Combinatorial Pattern Matching*. Springer, **1264**, pp. 262–274.

Schröck,E., du Manoir,S., Veldman,T., Schoell,B., Wienberg,J., Ferguson-Smith,M., Ning,Y., Ledbetter,D., Bar-Am,I., Soenksen,D., Garini,Y. and Ried,T. (1996) Multicolor spectral karyotyping of human chromosomes. *Science*, **273**, 494–497.

Smith,D., Huang,H. and Wang,L. (1998) Common fragile sites and cancer (review). *Int. J. Oncol.*, **12**, 187–196.

Tesler,G. (2002) GRIMM: genome rearrangements web server. *Bioinformatics*, **18**, 492–493.

Thompson,C. and Gray,J. (1993) Cytogenetic profiling using fluorescence *in situ* hybridization (FISH) and comparative genomic hybridization (CGH). *J. Cell Biochem. Suppl.*, **17G**, 139–143.

Volik,S., Zhao,S., Chin,K., Brebner,J., Herndon,D., Tao,Q., Kowbel,D., Huang,G., Lapuk,A. Kuo,W. *et al.* (2003) End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc. Natl Acad. Sci. USA*, **100**, (in press).

Wang,T., Rago,C., Silliman,N., Ptak,J., Markowitz,S., Willson,J., Parmigiani,G., Kinzler,K., Vogelstein,B. and Velculescu,V. (2002) Prevalence of somatic alterations in the colorectal cancer cell genome. *Proc. Natl Acad. Sci. USA*, **99**, 3076–3080.