nature biotechnology

ARTICLES

Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping

Paula J P de Vree[1,16], Elzo de Wit[1,2,16], Mehmet Yilmaz[2], Monique van de Heijning[2], Petra Klous[2], Marjon J A M Verstegen[1], Yi Wan[1], Hans Teunissen[1], Peter H L Krijger[1], Geert Geeven[1], Paul P Eijk[3], Daoud Sie[3], Bauke Ylstra[3], Lorette O M Hulsman[4], Marieke F van Dooren[4], Laura J C M van Zutven[4], Ans van den Ouweland[4], Sjef Verbeek[5,6], Ko Willems van Dijk[5,6], Marion Cornelissen[7], Atze T Das[7], Ben Berkhout[7], Birgit Sikkema-Raddatz[8], Eva van den Berg[8], Pieter van der Vlies[8], Desiree Weening[8], Johan T den Dunnen[9], Magdalena Matusiak[10,11], Mohamed Lamkanfi[10,11], Marjolijn J L Ligtenberg[12], Petra ter Brugge[13], Jos Jonkers[13], John A Foekens[14], John W Martens[14], Rob van der Luijt[15], Hans Kristian Ploos van Amstel[15], Max van Min[2], Erik Splinter[2] & Wouter de Laat[1,2]

Despite developments in targeted gene sequencing and whole-genome analysis techniques, the robust detection of all genetic variation, including structural variants, in and around genes of interest and in an allele-specific manner remains a challenge. Here we present targeted locus amplification (TLA), a strategy to selectively amplify and sequence entire genes on the basis of the crosslinking of physically proximal sequences. We show that, unlike other targeted re-sequencing methods, TLA works without detailed prior locus information, as one or a few primer pairs are sufficient for sequencing tens to hundreds of kilobases of surrounding DNA. This enables robust detection of single nucleotide variants, structural variants and gene fusions in clinically relevant genes, including *BRCA1* and *BRCA2*, and enables haplotyping. We show that TLA can also be used to uncover insertion sites and sequences of integrated transgenes and viruses. TLA therefore promises to be a useful method in genetic research and diagnostics when comprehensive or allele-specific genetic information is needed.

Current methods in genetic diagnostics and research are limited in their ability to uncover all possible genetic variation in genes of interest[1]. Clinical genetic tests, for example, often focus only on exons and therefore miss variants in the noncoding regulatory sequences (promoters and enhancers) of genes that can also disrupt gene function[2]. In addition, structural variants—such as copy number variants, translocations, insertions and inversions—are difficult to detect, particularly those that are balanced (i.e., inversions and translocations that are not accompanied by a loss or gain of sequences). Also, balanced structural variants can cause disease through the disruption of genes, creation of gene fusions or position effects—i.e., when a gene is placed under the control of different regulatory DNA elements[3,4]. Reliable detection of structural variants is hampered by the hypothesis-driven nature of current methods for targeted re-sequencing, in which the sequences to be analyzed are determined by the set of probes used in hybridization-based capture methods[5] or the primers in polymerase or ligase-based re-sequencing approaches[6]. Unknown sequences, such as those introduced by chromosomal rearrangements, are difficult to capture and re-sequence with these methods[7,8]. In addition, none of the existing targeted sequencing methods allow haplotyping, the allelic phasing of genetic variation. This information is useful, for example, when a person carries multiple recessive genetic variants that may coexist on one allele (giving carrier status) or be divided over both alleles (giving disease status).

In chromosome-conformation capture (3C)[9] and related methods such as 3C on chip or combined with sequencing[10,11] (4C, which enables searching of the genome for sequences contacting a site of interest), chromatin is crosslinked, fragmented and re-ligated to identify, on the basis of their ligation efficiency, genomic loci that are in close spatial proximity in the nucleus. Although these methods are used to detect genome folding inside cells, the resulting contact profiles typically show high enrichment of sequences directly neighboring

[1]Hubrecht Institute-KNAW and University Medical Center Utrecht, Utrecht, the Netherlands. [2]Cergentis B.V., Utrecht, the Netherlands. [3]Department of Pathology, VU University Medical Center, Amsterdam, the Netherlands. [4]Department of Clinical Genetics, Erasmus Medical Center, Rotterdam, the Netherlands. [5]Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. [6]Department of Endocrinology, Leiden University Medical Center, Leiden, the Netherlands. [7]Laboratory of Experimental Virology, Department of Medical Microbiology, Center for Infection and Immunity Amsterdam (CINIMA), Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands. [8]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [9]Leiden Genome Technology Center, Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands. [10]Department of Medical Protein Research, VIB, Ghent, Belgium. [11]Department of Biochemistry, Ghent University, Ghent, Belgium. [12]Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands. [13]Division of Molecular Pathology and Cancer Genomics Center, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [14]Department of Medical Oncology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, the Netherlands. [15]Department of Medical Genetics, University Medical Center Utrecht, Utrecht, the Netherlands. [16]These authors contributed equally to this work. Correspondence should be addressed to E.S. (erik.splinter@cergentis.com) or W.d.L. (w.delaat@hubrecht.eu).

Received 26 November 2013; accepted 16 June 2014; published online 17 August 2014; doi:10.1038/nbt.2959

a segment of interest as predicted by polymer physics theory[12]. This enrichment decreases with increased site separation on the linear chromosome[10,13]. In studies seeking to understand the three-dimensional (3D) properties of chromosomes, minor deviations from this pattern are the signals of interest. However, this same pattern has been used to show that 4C can identify chromosomal rearrangement partners[14,15] and that Hi-C technology, a 3C variant that uncovers contacts throughout the genome, can be used for chromosome-wide scaffolding[16,17]. We have used the principles of proximity ligation to develop TLA, a strategy for targeted re-sequencing. The method enables re-sequencing and haplotyping of genes or genomic regions, of sizes ranging from tens to hundreds of kilobases, and uncovers a wide range of genetic changes, including single nucleotide variants (SNVs) and structural variants.

RESULTS

TLA

For targeted sequencing, we wished to capture and analyze the DNA segments that surround a selected site (the 'anchor') in a region of interest (**Fig. 1**). 4C does not allow this because crosslinked and ligated DNA fragments are trimmed to selectively amplify and sequence only their ends (**Fig. 1b** and **Supplementary Fig. 1**). We therefore developed a procedure to ensure the amplification of neighboring fragments without loss of sequence and the retrieval of maximum sequence information per crosslinked DNA molecule. Both TLA and 4C (compared in **Supplementary Fig. 1**) involve digestion of the crosslinked chromatin with a restriction enzyme (here, NlaIII) that recognizes a 4-base-pair site (4-cutter) followed by ligation to obtain large DNA circles containing multiple crosslinked NlaIII restriction fragments (**Fig. 1a**, **i–v**). After de-crosslinking, in TLA the circles are further digested with a secondary restriction enzyme, NspI, a '5-cutter' that creates DNA fragments with an average size of 2 kilobases (kb) carrying multiple ligated NlaIII fragments. NspI shares a core recognition sequence (CATG) with NlaIII but has additional sequence requirements, such that NlaIII fragments are not are lost from sequence analysis (**Supplementary Fig. 1a**). After digestion with NspI, molecules are circularized and amplified by PCR using anchor-specific, outward-oriented primers (**Fig. 1a**, **vi–vii**). The molecules containing the anchor will be fused to many different locus-derived NlaIII fragments, so PCR with a single primer set at the anchor will result in the amplification of many NlaIII fragments across tens to hundreds of kilobases of surrounding DNA (**Fig. 1a**, **vi–viii**). The PCR-amplified material is then sonicated and adaptor ligated for high-throughput sequencing. Mapping the sequenced reads to a reference genome allows building of contigs representing the sequence of a genetic locus of interest.

TLA applied to the *BRCA* genes

To validate TLA we first focused on the *BRCA1* gene, which is implicated in hereditary breast and ovarian cancer and in the response to certain drugs to treat breast cancer[18]. PCR-based exon sequencing methods are often used for *BRCA1* testing, and these genetic tests typically involve using at least 30 amplicons (but often many more) to analyze only the *BRCA1* exons. For TLA in human leukemia K562 cells, we used a single anchor primer pair inside the gene body to capture and amplify sequences across the 81-kb *BRCA1* gene (**Fig. 2a**,**b** and **Supplementary Fig. 2**). As we expected on the basis of chromosome-folding properties, sequence coverage was high around the anchor, decayed with increased distance on the *cis* chromosome and was sparse on all other chromosomes. Importantly, approximately 30–40% of sequences retrieved were derived from within the *BRCA1* or *BRCA2* genes (**Supplementary Fig. 3**). Within the target gene, coverage was highly nonuniform (**Fig. 2a**,**b** and **Supplementary Fig. 4**), as sequences immediately flanking the anchor primers were exceptionally abundant. This is due largely to undigested DNA (as revealed by many reads across endogenous NlaIII restriction sites; data not shown); incomplete enzymatic digestion of crosslinked chromatin is a known feature of 3C-based protocols, for which digestion efficiencies are typically 50–80%[11]. Despite this excess of retrieved sequences flanking the anchors, we found that 150 Mb of sequence

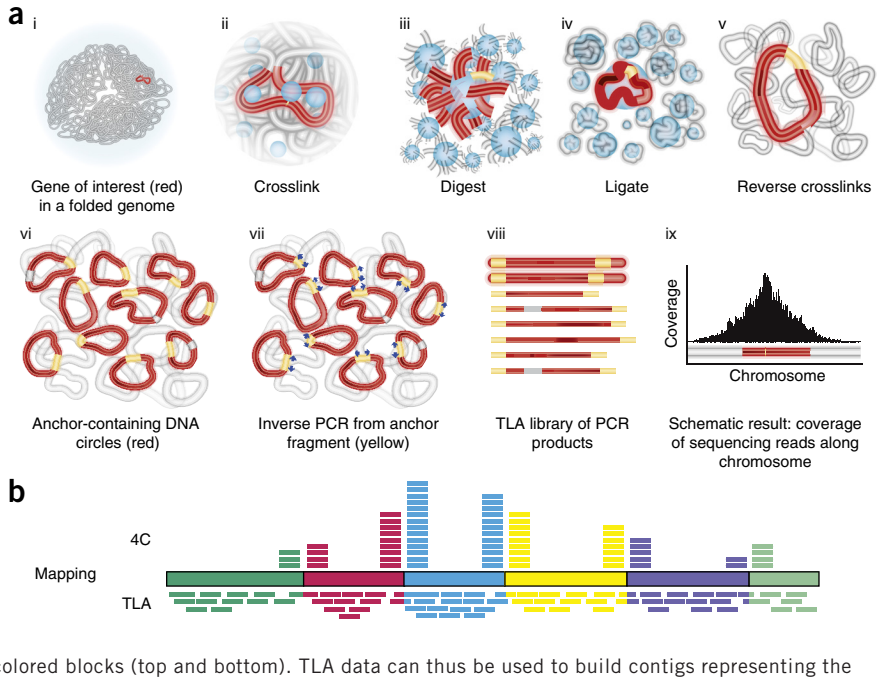


Figure 1 Targeted sequencing using TLA. (**a**) Neighboring sequences that form a gene or genetic locus (red) are in close spatial proximity (i) and therefore are preferentially crosslinked (ii). Digestion with a frequently cutting enzyme (iii) and ligation (iv) results in large DNA circles composed of multiple crosslinked restriction fragments (v). Different copies of a locus (from different cells) result in DNA circles composed of co-captured restriction fragments. Limited trimming (with a compatible but less frequently cutting enzyme) and ligation creates PCR-amplifiable DNA circles (vi). Fragments captured with a fragment of interest (the anchor sequence, yellow) are selectively PCR-amplified with anchor-specific inverse PCR primers (blue arrows) (vii). The resulting sample (viii) is highly enriched for locus-specific sequences and can be processed with standard library preparation procedures for next-generation sequencing. Mapped reads originate from the locus of interest and collectively span tens of kilobases (ix). (**b**) In TLA, the entire restriction fragments are amplified and sequenced, whereas 4C analyzes the ends of the fragments. Sequencing reads are shown as colored blocks (top and bottom). TLA data can thus be used to build contigs representing the sequence of a genetic locus of interest.

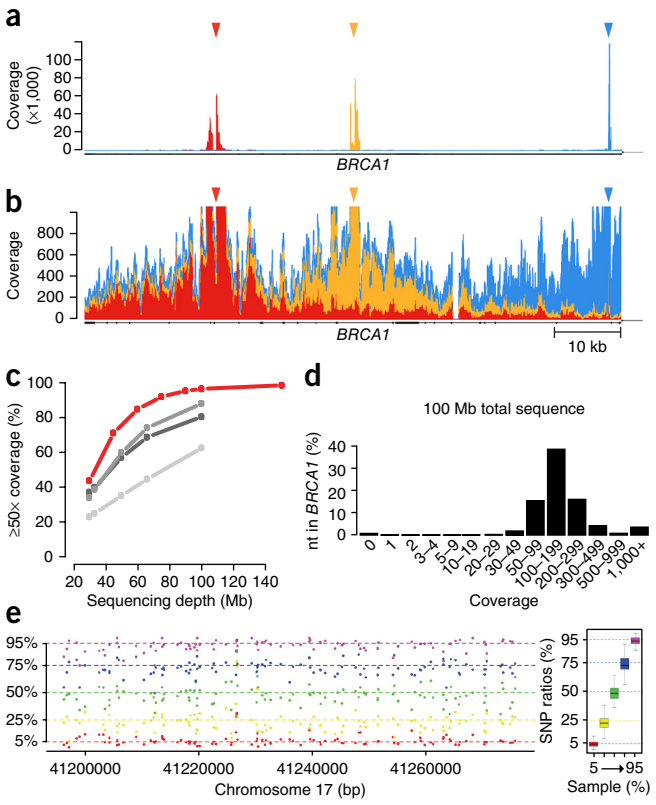


Figure 2 TLA analysis of the *BRCA1* gene. (**a**) Sequencing coverage across the 81-kb *BRCA1* gene shows that sequences immediately neighboring the three TLA anchors (arrowheads) are highly overrepresented; this compromises coverage equality but enables haplotyping and facilitates rearrangement detection. (**b**) Trimming the extreme signals around anchors reveals high coverage (running median with a 101-nt window) across the rest of *BRCA1* gene (with the exception of an intronic, highly AT-rich gap that is also difficult to sequence by whole-genome sequencing)[19]. Coverage for three anchors (red, orange and blue arrowheads), each obtained from 100-Mb sequence depth. (**c**) Required sequencing depth (*x* axis) for three individual anchors (gray) and their combination (red) to reach ≥50-fold coverage across a given percentage of the *BRCA1* gene (*y* axis). (**d**) Histogram showing, at 100-Mb sequencing depth of the combined anchors, the percentage of all *BRCA1* nucleotides analyzed at a given (binned) coverage. (**e**) Two cell lines homozygous for *BRCA1* but with distinct haplotypes were mixed in different ratios (95:5 (magenta), 75:25 (blue), 50:50 (green), 25:75 (yellow), 95:5 (red)) before TLA processing. Left, each dot represents a variant position in the *BRCA1* gene and shows the percentage of reads originating from a given haplotype, as determined by TLA; dashed lines show the expected percentage for each mixture. Right, box and whisker plots (boxes, median and quartiles; whiskers, 5th and 95th percentiles) summarizing the measured ratios for all variants for each sample. nt, nucleotides.

data (~3% of an Illumina MiSeq run) was required to provide least 50-fold coverage across >98% of nucleotides in the *BRCA1* locus (**Fig. 2c**,**d**). This implies that each nucleotide of interest is enriched 1,000-fold or more.

For reliable base calling we required a minimum of 50-fold coverage and a strand bias no higher than 90%. To exclude the possibility that coverage in TLA is predominantly obtained by the amplification of single crosslinked fragments (single captures), we applied TLA to titration mixes of two cell lines (SUM149PT and MDA-MB-436) that are homozygous for *BRCA1*. Single-nucleotide differences between the cell lines were accurately identified in the correct ratios across the entire locus, even in mixes with a 95:5 allele imbalance. This was achieved when we used 800 ng template (**Fig. 2e**) and when input DNA was reduced to 200 ng (**Supplementary Fig. 5**). This demonstrates that a sufficient number of independent alleles were analyzed for reliable SNP calling and suggests that TLA would be able to identify variants in subpopulations of tumor cells.

Having tested the method on K562 cells and on the cell mixtures, we applied TLA at the *BRCA1* and *BCRA2* genes to 20 anonymized biopsies of breast tumors and 11 anonymized blood samples (we did a simultaneous analysis of the two *BRCA* genes except when informed consent limited analysis to only one of the genes). We typically started with 10 million cells, but also obtained good results with 1 million cells, as the PCR step uses only a fraction of DNA (800 ng is the content of $<1.4 \times 10^5$ cells).

For all of the samples tested, at least 99% of the exonic and 98% of the total gene sequences reached the aforementioned quality thresholds. There is little bias against large or small NlaIII fragments (**Supplementary Fig. 6**) because the amplification step occurs on large DNA circles composed of multiple, often partially undigested, NlaIII fragments. However, large NlaIII fragments containing long stretches of low-complexity sequences had lower sequencing coverage. In *BRCA1*, for example, one highly AT-rich region of ~1 kb seemed difficult to amplify (**Fig. 2b**); this region also shows reduced coverage in the 1000 Genomes Project data[19]. Long reads (150–250 nucleotides) can be used to map repeat sequences, as such reads often include unique sequences (**Supplementary Fig. 7**).

From the analysis of all samples, 7 uncommon SNVs (SNVs not found in the dbSNP database[20] (build 138)), 11 insertions or deletions (indels) and 193 common SNVs were identified and validated by Sanger sequencing (**Supplementary Table 1**). Three variants could not be verified, of which one lay in a repetitive long interspersed nuclear element. To compare the performance of TLA to whole-genome sequencing (WGS), we applied it to the analysis of *BRCA1* and *BRCA2* in the deep-sequenced GM12878 cell line (European Nucleotide Archive ERP001775). Of the 160.2 kb of *BRCA1* and *BRCA2* sequence retrieved by TLA, only one base pair was called differently between TLA and WGS.

Haplotyping of genomic loci

Homologous chromosomes occupy physically distinct territories in human somatic cells[21]. Consequently, there is a high probability that co-captured intrachromosomal fragments originate from the same parental chromosome. This principle has been used to reconstruct, at low resolution, the haplotypes of complete chromosomes by the Hi-C approach[22]. We reasoned that when paired-end sequencing is applied to TLA libraries, the excess of sequences flanking the anchor retrieved should allow haplotyping when a discriminating SNV is in or near the anchor fragment and is amplified along with the anchor in the PCR step (**Fig. 3a**). We determined that 5–10% of all mappable read pairs contained such a SNV in one of the read ends (**Fig. 3a**), which allowed the phasing of SNVs identified in the paired end read ends.

We designed anchor primers near four allele-specific SNVs in the *BRCA1* gene. Each anchor could phase 77–90% of the 103 heterozygous SNVs in the 81-kb *BRCA1* gene. When used together they were capable of unambiguously assigning 101 of the 103 SNVs (>98%) to a *BRCA1* allele (**Fig. 3b**,**c**). We obtained a similar percentage of allelically assigned SNVs when we applied this strategy to two other genes (93% and 97% in *BRCA2* and *ACADM*, respectively) (**Supplementary Fig. 8**). We then sampled the sequencing data to understand the sequencing depth required for haplotyping over distance intervals. Even with a relatively low sequencing amount (40 Mb), SNVs >150 kb away from the anchor (in a 300-kb interval around the anchor) could still be haplotyped (**Fig. 3d** and **Supplementary Fig. 8**).

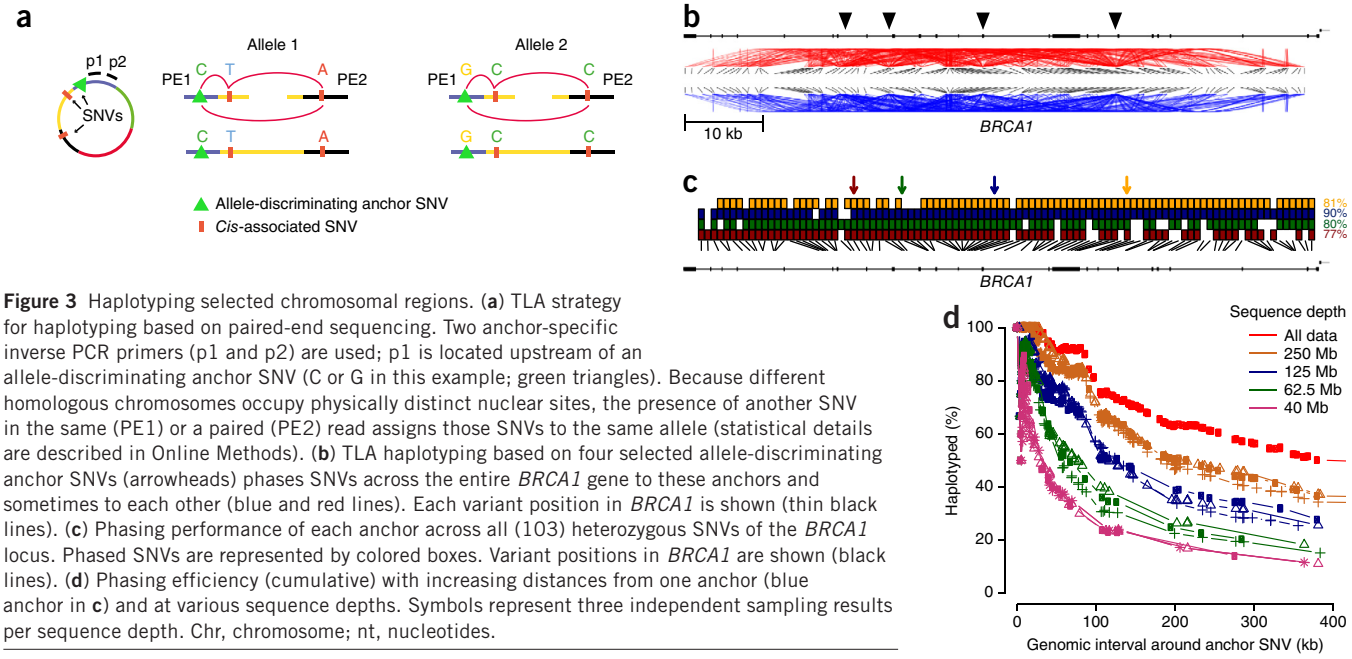
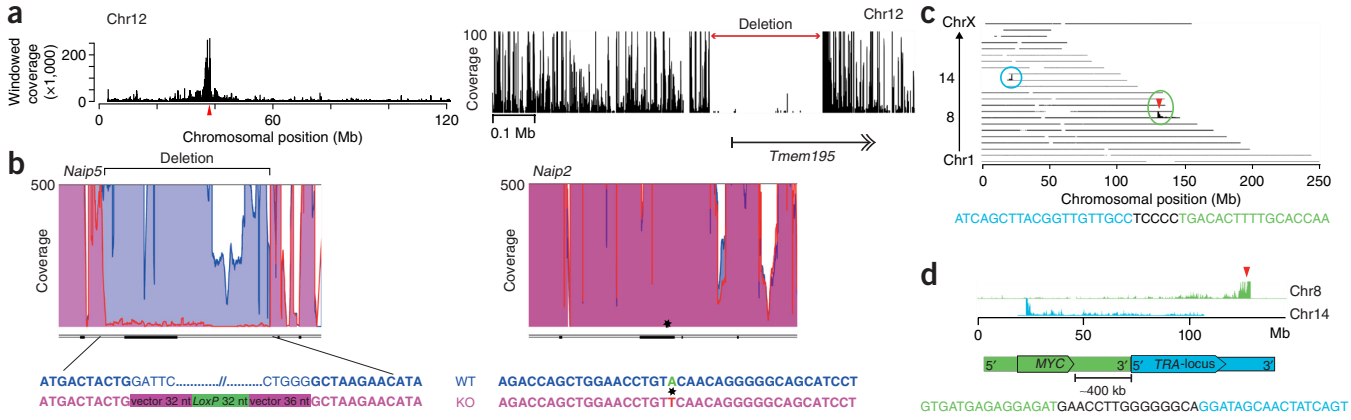


Figure 3 Haplotyping selected chromosomal regions. (**a**) TLA strategy for haplotyping based on paired-end sequencing. Two anchor-specific inverse PCR primers (p1 and p2) are used; p1 is located upstream of an allele-discriminating anchor SNV (C or G in this example; green triangles). Because different homologous chromosomes occupy physically distinct nuclear sites, the presence of another SNV in the same (PE1) or a paired (PE2) read assigns those SNVs to the same allele (statistical details are described in Online Methods). (**b**) TLA haplotyping based on four selected allele-discriminating anchor SNVs (arrowheads) phases SNVs across the entire *BRCA1* gene to these anchors and sometimes to each other (blue and red lines). Each variant position in *BRCA1* is shown (thin black lines). (**c**) Phasing performance of each anchor across all (103) heterozygous SNVs of the *BRCA1* locus. Phased SNVs are represented by colored boxes. Variant positions in *BRCA1* are shown (black lines). (**d**) Phasing efficiency (cumulative) with increasing distances from one anchor (blue anchor in **c**) and at various sequence depths. Symbols represent three independent sampling results per sequence depth. Chr, chromosome; nt, nucleotides.

Sequencing transgenes and their integration sites

Because TLA re-sequences parts of the genome on the basis of their physical proximity to an anchor fragment, it should be suitable for mapping transgenes and characterizing their genomic integration sites. Currently, such mapping is often done with strategies such as primer extension, but these have limitations—when the exact size of the integrated cassette is unknown, for example. We used TLA to study a previously unmappable transgene[23] encoding a variant of apolipoprotein E (ApoE) in spleen cells from a transgenic mouse and found that it was located in the *Agmo* gene (also known as *Tmem195*), where it had deleted part of the endogenous gene (**Fig. 4a**). This deletion might explain our observation that these transgenic mice



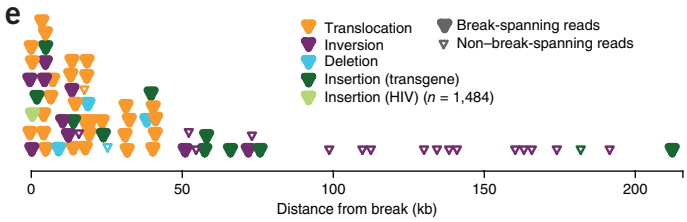


Figure 4 TLA applied to transgenes and chromosomal rearrangements. (**a**) Analysis using TLA of a transgenic mouse line with a randomly integrated variant of the gene encoding ApoE[23]. Transgenic sequence information was used as the anchor. The plot (left) shows that the transgene integrated in chromosome 12 around position 38 Mb (arrowhead). The close-up (right) shows integration has deleted part of the endogenous *Tmem195* gene. (**b**) TLA analysis shows Cre-mediated deletion of nearly 7 kb across an exon in *Naip5* knockout mice (KO, magenta) but not in wild-type mice (WT, blue). A single *loxP* site (green) and some flanking vector sequences remain at the site of deletion. Exon positions are indicated as black boxes below graph (left). A missense mutation was also found ~65 kb away in the neighboring *Naip2* gene (star, right). (**c**,**d**) Detection of chromosomal rearrangements with TLA. The human acute lymphoid leukemia (ALL) cell line MOLT-16 (ref. 32) carries a translocation between the *TRA-TRD* locus on chromosome 14 and the *MYC*-flanking gene desert on chromosome 8. (**c**) Genome-wide TLA coverage plot shows that a TLA anchor on chromosome 8 (red arrowhead) identifies *TRA-TRD* on chromosome 14 (blue circle). (**d**) An anchor (red arrowhead) on the other side of the breakpoint on chromosome 8 shows how *MYC* is fused to the TRA locus. The sequence of the breakpoint is given. Plots show sum of the coverage in 10-kb window, in percentiles (trimmed at 98th percentile). (**e**) A summary of 71 independent TLA experiments showing whether nucleotide resolution of the breakpoint could be obtained (filled triangles, yes; open triangles, no) for anchors at increasing distance from the breakpoint. Chr, chromosome.

could not be bred to homozygosity. We also tested the capability of TLA to uncover viral integration sites and identified 1,484 HIV genomic insertions at base-pair resolution in SupT1 cells infected with HIV. Such mapping is usually done by PCR or ligation-mediated PCR[24,25], but as TLA re-sequences the integrated viral genome, it can also be used to discover the HIV type most prevalent in the cells[26] (**Supplementary Fig. 9**).

We then investigated whether TLA could provide information about locus-specific integration events by applying the method to *Naip5* knockout mice[27]. Unambiguous confirmation of the correct disruption of *Naip5* had not been possible using other techniques, including Southern blotting, as the region is highly repetitive and contains multiple other *Naip* genes that share up to 94% sequence identity with *Naip5*. TLA confirmed correct integration of the targeting cassette in *Naip5*, showed that the nearby homologous *Naip1* and *Naip6* genes were intact and revealed the *in cis* coselection of a previously undetected SNV causing a tyrosine-to-asparagine substitution in the protein encoded by *Naip2* (**Fig. 4b**). The coselection of passenger mutations has been linked to mouse phenotypes[28].

Identification of structural variants

Structural rearrangements, by definition, introduce unknown sequences into a locus of interest and are therefore difficult to detect by targeted re-sequencing methods that depend on sequence complementarity (such as PCR and hybridization capture). In TLA, unknown sequences introduced through rearrangements should be captured and sequenced with the same efficiency as locus-intrinsic sequences. Moreover, even if there are no breakpoint-spanning reads, visual inspection of TLA sequencing reads plotted along the chromosomes may still be sufficient to uncover rearrangements. To test this, we carried out >50 independent TLA experiments on human cell lines known to carry chromosomal rearrangements (deletions, translocations and inversions) but for which there is no exact breakpoint information (**Supplementary Table 2**). Sequencing coverage plots revealed the large rearrangements, even for breakpoints >180 kb away from the anchor sequence (**Fig. 4c–e**). In such cases, coverage distribution can predict the type of chromosomal rearrangement: for example, translocations show a peak of high coverage on an unrelated chromosome (**Fig. 4c,d**); large deletions show a drop and rise in coverage, respectively, inside and beyond the deletion; and inversions (>1 Mb) show a coverage pattern that goes up instead of down with greater site separation on the reference genome (**Supplementary Fig. 10**). The detection of smaller and/or more complex rearrangements requires closer data inspection; although the coverage plots can give some information, the exact type of rearrangement can often be resolved unambiguously only by sequences spanning the breakpoints. In 31 of 33 experiments with anchors mapping within 20 kb of the break site, we were able to identify such reads and to map breakpoints at base-pair resolution; this often required $<1 \times 10^6$ sequencing reads (**Fig. 4e** and **Supplementary Fig. 11**). Beyond this chromosomal distance, rearrangements were still detectable from TLA profiles, but break-spanning reads were not always found (**Fig. 4e** and **Supplementary Table 2**). In such instances, breakpoint mapping at base-pair resolution, if required, could be accomplished by using an anchor primer pair closer to the breakpoint. Information is also obtained that phases the variants on the chromosomes involved in rearrangements (**Supplementary Fig. 12**), which would, for example, enable discernment of which alleles contribute to fusion genes.

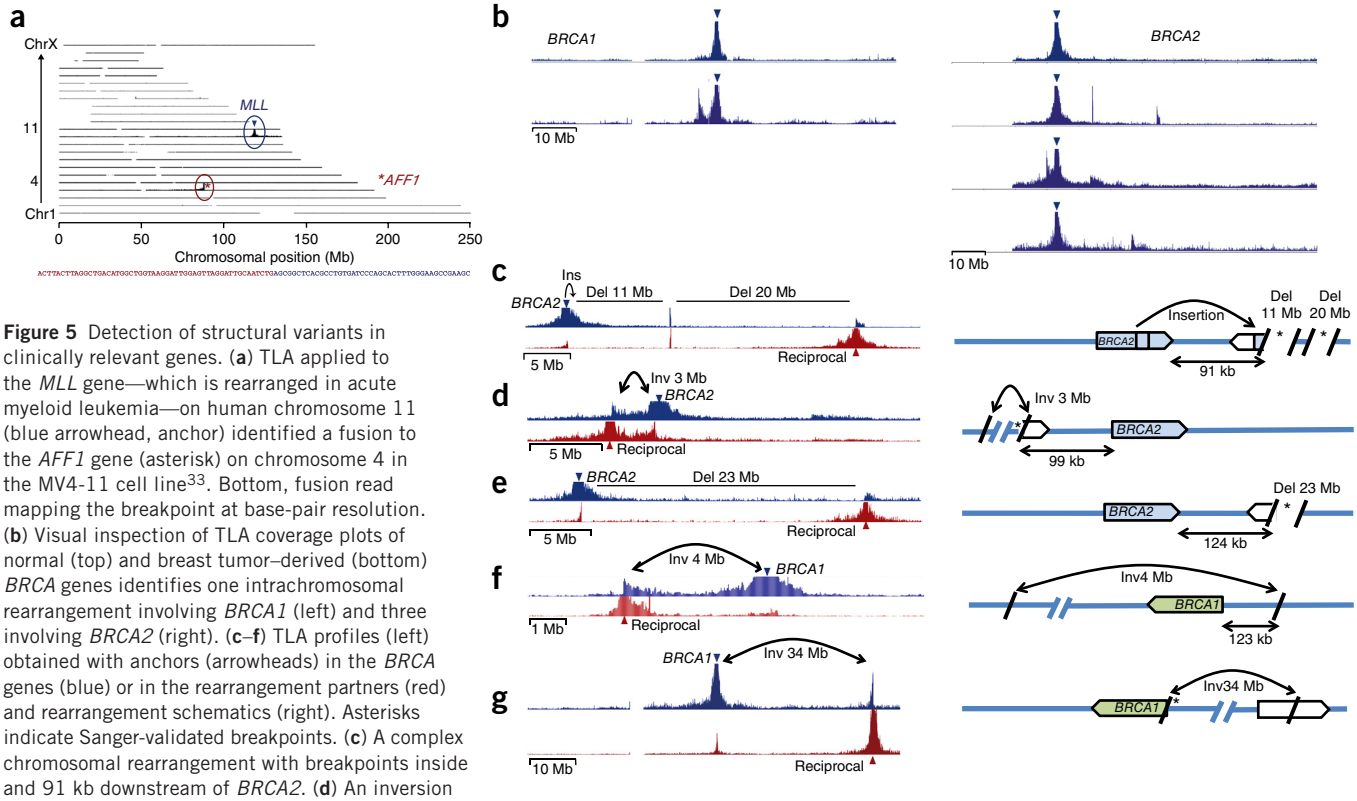


Figure 5 Detection of structural variants in clinically relevant genes. (**a**) TLA applied to the *MLL* gene—which is rearranged in acute myeloid leukemia—on human chromosome 11 (blue arrowhead, anchor) identified a fusion to the *AFF1* gene (asterisk) on chromosome 4 in the MV4-11 cell line[33]. Bottom, fusion read mapping the breakpoint at base-pair resolution. (**b**) Visual inspection of TLA coverage plots of normal (top) and breast tumor–derived (bottom) *BRCA* genes identifies one intrachromosomal rearrangement involving *BRCA1* (left) and three involving *BRCA2* (right). (**c**–**f**) TLA profiles (left) obtained with anchors (arrowheads) in the *BRCA* genes (blue) or in the rearrangement partners (red) and rearrangement schematics (right). Asterisks indicate Sanger-validated breakpoints. (**c**) A complex chromosomal rearrangement with breakpoints inside and 91 kb downstream of *BRCA2*. (**d**) An inversion with breakpoint 99 kb upstream of *BRCA2*. (**e**) A large deletion with a breakpoint 124 kb downstream of the *BRCA2* gene. (**f**) A complex inversion with a breakpoint 123 kb upstream of *BRCA1*. (**g**) Application of TLA to a breast cancer xenograft identifies a >30-Mb inverted chromosomal part that creates a previously unidentified fusion of the genes *SEPT9* and *BRCA1*. Breakpoints validated by Sanger sequencing are shown (arrowheads). *y* axis (**b**–**g**) shows sum of the coverage in 10-kb window in percentiles (trimmed at 98th percentile). Del, deletion; inv, inversion; ins, insertion.

Structural variants and gene fusions in cancer samples

Finally, we used TLA to sequence clinically relevant genes and search for structural variants in or near the coding sequences. We first focused on the *MLL* gene, which can fuse to diverse genomic partners to drive leukemia[29]. Determining specific *MLL* rearrangements has prognostic value and is routinely done by karyotyping or fluorescence *in situ* hybridization (FISH). However, these approaches have low resolution, and FISH can test for only one previously known fusion partner at a time. Using TLA we mapped eight chromosomal breakpoints involving four different rearrangement partners of *MLL* at base-pair resolution in the four leukemic cell lines tested (**Fig. 5a** and **Supplementary Fig. 13**). Breakpoint information is useful for the design of specific PCR tests that can be used in the clinical context to quantify minimal residual disease[30].

We next asked whether structural rearrangements such as chromosomal inversions or translocations also occur in *BRCA1* and *BRCA2*. SNVs, small indels and copy number variants have been found in these genes, but, to our knowledge, inversions and translocations have not. We analyzed tumor samples from 18 patients with early-onset breast cancer for which standard methodology had failed to detect germline inherited deleterious *BRCA1* or *BRCA2* mutations. We found inversion or complex rearrangements of *BRCA1* or *BRCA2* in four of these cases (**Fig. 5b** and **Supplementary Fig. 14**). In three cases, the breakpoints were located <125 kb away from the *BRCA1* or *BRCA2* gene, and in one instance the rearrangement involved breaks inside as well as downstream of *BRCA2* (**Fig. 5c–f**). In the one case for which matching blood was available (**Fig. 5c**), the blood sample did not carry the rearrangement, suggesting that it is a somatic rearrangement. For the other cases, no nontumorous material was available, so we could not verify whether the rearrangements were germline or somatic. Nevertheless, our recurrent finding of chromosomal rearrangements near the *BRCA* genes in early-onset breast tumors was intriguing. We also applied TLA to three xenografts[31] from breast tumors that originally carried methylated (and therefore inactivated) *BRCA1*, but that, because of a gained drug resistance, were suspected to have re-activated *BRCA1* alleles. In all these cases we identified large chromosomal inversions creating gene fusions that placed the *BRCA1* coding region under a different promoter (**Fig. 5g** and data not shown). These findings demonstrate that *BRCA1* and *BRCA2* function can be affected by large structural changes that are difficult to detect by standard methods of analysis.

DISCUSSION

Our data show that TLA can be used to sequence genes and/or other chromosomal regions of interest with little requirement for prior sequence knowledge. This enables simultaneous detection of large balanced and unbalanced chromosomal rearrangements as well as SNVs and indels and allows extensive characterization of transgene integration sites and haplotyping across large genomic intervals. The procedure can be multiplexed and is amenable to automation, enabling analysis of large and/or multiple genes.

The current TLA protocol requires cells as the input, limiting its application to cell lines and blood or tissue samples. If *in vitro* chromatin assembly becomes possible, TLA could become compatible with purified genomic DNA; alternatively, other strategies involving crosslinking, fragmenting and re-ligation for targeted sequencing of genes of interest are possible. One such future application of TLA might be for the analysis of formalin-fixed paraffin embedded material (the form in which most surgically removed tissue or tumor biopsies are stored). As this type of preparation involves formaldehyde fixation, in cases where the DNA is not entirely degraded the samples might be suitable for targeted re-sequencing by modified TLA procedures. In summary, TLA provides the opportunity for reliable detection of both SNVs and structural variants in selected genomic regions and therefore promises to be a useful tool in clinical diagnostics and research in all cases requiring complete sequence information in and around genes.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequence data have been deposited to European Nucleotide Archive under accession number ERP005535.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank T.P. Driessen for figure graphics. This work was supported by the Netherlands Genomics Initiative (NGI) pre-seed grants 93608003 and 93611010 and a proof of concept grant from the Cancer Genomics Center (CGC) to W.d.L. and an Innovation Credit from NL Agency to Cergentis. P.J.P.d.V. is supported by a Dutch Cancer Foundation grant KWF (2009-4459 to J.A.F., J.W.M. and W.d.L.).

AUTHOR CONTRIBUTIONS

P.J.P.d.V., E.d.W., M.v.M., E.S. and W.d.L. conceived the experiments and analyzed the data; P.J.P.d.V., M.Y., M.v.d.H., P.K., M.J.A.M.V., Y.W., H.T., P.H.L.K., G.G. and E.S. performed the experiments and analyzed the data; M.v.M. and W.d.L. invented TLA. All other authors provided patient samples and analyzed data. E.S., E.d.W. and W.d.L. wrote the manuscript with input from P.J.P.d.V., P.H.L.K., G.G., D.S., B.Y., J.J., L.J.C.M.v.Z., M.L., M.C., B.S.-R., K.W.v.D., M.J.L.L., H.K.P.v.A., J.T.d.D., J.W.M. and M.v.M.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Katsanis, S.H. & Katsanis, N. Molecular genetic testing and the future of clinical genomics. *Nat. Rev. Genet.* **14**, 415–426 (2013).
2. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
3. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J.O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
4. Kleinjan, D.A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8–32 (2005).
5. Harakalova, M. *et al.* Multiplexed array-based and in-solution genomic enrichment for flexible and cost-effective targeted next-generation sequencing. *Nat. Protoc.* **6**, 1870–1886 (2011).
6. Frank, T.S. *et al.* Sequence analysis of *BRCA1* and *BRCA2*: correlation of mutations with family history and ovarian cancer risk. *J. Clin. Oncol.* **16**, 2417–2425 (1998).
7. Altmüller, J., Budde, B.S. & Nurnberg, P. Enrichment of target sequences for next generation sequencing applications in research and diagnostics. *Biol. Chem.* **395**, 231–237 (2014).
8. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111–118 (2010).
9. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
10. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).
11. van de Werken, H.J. *et al.* 4C technology: protocols and data analysis. *Methods Enzymol.* **513**, 89–112 (2012).
12. Rippe, K. Making contacts on a nucleic acid polymer. *Trends Biochem. Sci.* **26**, 733–740 (2001).
13. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
14. Simonis, M. *et al.* High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology. *Nat. Methods* **6**, 837–842 (2009).
15. Homminga, I. *et al.* Integrated transcript and genome analyses reveal NKX2–1 and MEF2C as potential oncogenes in T cell acute lymphoblastic leukemia. *Cancer Cell* **19**, 484–497 (2011).

16. Burton, J.N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
17. Kaplan, N. & Dekker, J. High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nat. Biotechnol.* **31**, 1143–1147 (2013).
18. Fong, P.C. *et al.* Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *N. Engl. J. Med.* **361**, 123–134 (2009).
19. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
20. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
21. Bolzer, A. *et al.* Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* **3**, e157 (2005).
22. Selvaraj, S., Dixon, J.R., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31**, 1111–1118 (2013).
23. van den Maagdenberg, A.M. *et al.* Transgenic mice carrying the apolipoprotein E3-Leiden gene exhibit hyperlipoproteinemia. *J. Biol. Chem.* **268**, 10540–10545 (1993).
24. Uren, A.G. *et al.* A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nat. Protoc.* **4**, 789–798 (2009).
25. Koudijs, M.J. *et al.* High-throughput semiquantitative analysis of insertional mutations in heterogeneous tumors. *Genome Res.* **21**, 2181–2189 (2011).
26. Smith, S.D. *et al.* Clinical and biologic characterization of T-cell neoplasias with rearrangements of chromosome 7 band q34. *Blood* **71**, 395–402 (1988).
27. Lightfield, K.L. *et al.* Critical function for Naip5 in inflammasome activation by a conserved carboxy-terminal domain of flagellin. *Nat. Immunol.* **9**, 1171–1178 (2008).
28. Kayagaki, N. *et al.* Non-canonical inflammasome activation targets caspase-11. *Nature* **479**, 117–121 (2011).
29. Meyer, C. *et al.* The MLL recombinome of acute leukemias in 2013. *Leukemia* **27**, 2165–2176 (2013).
30. Leary, R.J. *et al.* Development of personalized tumor biomarkers using massively parallel sequencing. *Sci. Transl. Med.* **2**, 20ra14 (2010).
31. Evers, B. *et al.* A high-throughput pharmaceutical screen identifies compounds with specific toxicity against BRCA2-deficient tumors. *Clinical cancer research: an official journal of the American Association for Cancer Research* **16**, 99–108 (2010).
32. McKeithan, T.W. *et al.* Molecular cloning of the breakpoint junction of a human chromosomal 8;14 translocation involving the T-cell receptor α-chain gene and sequences on the 3′ side of MYC. *Proc. Natl. Acad. Sci. USA* **83**, 6636–6640 (1986).
33. Lange, B. *et al.* Growth factor requirements of childhood acute leukemia: establishment of GM-CSF-dependent cell lines. *Blood* **70**, 192–199 (1987).

ONLINE METHODS

Sample collection. K562, MDA-MB-436, Molt16 and SUM149PT cells as well as immortalized cells obtained from a patient with complex chromosomal rearrangements (CCR)[34] were available in our institutes, periodically tested for mycoplasma and cultured according to standard culturing procedures. The adherent cell lines MDA-MB-436 and SUM149PT were detached and made into single cell suspensions using a 40-µm cell strainer before starting the TLA procedure. PBMCs from whole blood were isolated via Ficoll separation. For TLA analysis on breast tumor tissue from hereditary breast cancer cases in which exonic mutations were excluded by routine diagnostics[35], retrospectively collected 100,000-g pellets containing crude nuclear extracts were used as starting material. These 100,000-g crude nuclear extracts remained after sample preparation for estrogen receptor and progesterone receptor measurements using enzyme immunoassays as part of the routine diagnostics before ER immunohistochemistry became the standard[36]. These crude nuclear extracts have been stored in liquid nitrogen (LN2) until use in the current study for TLA analysis. Xenograft tissue samples were snap frozen in LN2, then a Retsch MM400 was used for cryogenic grinding. After dissection, mouse spleen tissue was made into a single cell suspension by gently pressing it through a 40-µm cell strainer. Contaminating red blood cells were lysed by incubation in RBC lysis buffer (154 mM NH_4Cl; 10 mM $KHCO_3$; 200 µM EDTA) before starting the TLA procedure. When required, isolated samples were frozen in 10% DMSO/, 10% FBS and DMEM for storage at −80 °C.

For the mixing of cell lines to determine TLA sensitivity, 50 µl of harvested cell suspensions from SUM149PT and MDA-MB-436 were applied on a CASY counting machine for cell counting (InnoVartis; Roche). On the basis of data from CASY counting, cell lines were mixed in the following proportions: SUM149PT 5% (MDA-MB-436 95%); SUM149PT 25% (MDA-MB-436 75%); SUM149PT 50% (MDA-MB-436 50%); SUM149PT 75% (MDA-MB-436 25%); SUM149PT 95% (MDA-MB-436 5%). Mixed samples had 1×10^7 cells as starting material for TLA preparation.

For the generation of the virus stock, the HIV-1LAI plasmid was transfected into SupT1 cells[26] by electroporation. SupT1 cells were cultured in RPMI advanced medium with 1% heat-inactivated FBS (FBS) and penicillin, streptomycin and L-glutamine. Viral spread was monitored by syncytium formation, and the virus-containing culture supernatant was collected at the peak of infection. 3×10^7 SupT1 cells were infected with this virus stock (the equivalent of 85 ng CA-p24) in 30 ml medium and cultured for 5 d. We used 1×10^7 cells (with the HIV-1 genome integrated in the host cell genome) for TLA.

Human subjects provided informed consent to the use of material for the genetic testing described in accordance with the rules of the Institutional Review Board of the UMC Utrecht and the Erasmus Medical Center Rotterdam.

TLA sample preparation. A detailed step-by-step protocol is provided in **Supplementary Protocol**. In brief, the initial steps of the TLA procedure are performed as previously described[10,37]. Cells are crosslinked using formaldehyde, then DNA is digested with NlaIII. The sample is then ligated, crosslinks are reversed and the DNA is purified. To obtain circular chimeric DNA molecules for PCR amplification, the DNA molecules were trimmed with NspI and ligated at a DNA concentration of 5 ng/µl to promote intramolecular ligation. Importantly, NspI was chosen for its RCATGY recognition sequence that encompasses the CATG recognition sequence of NlaIII. As a consequence, only a subset of NlaIII (CATG) sites were (re-)digested, generating DNA fragments of approximately 2 kb and allowing the amplification of entire restriction fragments. After ligation the DNA was purified, and eight 25-µl PCR reactions, each containing 100 ng template, were pooled for sequencing. Sequences of the inverse primers, which were designed using Primer3 software[38], can be found in **Supplementary Table 2**. Illumina NGS library preparations were performed according to manufacturer's protocols. We performed sequencing of TLA libraries on MiSeq, HiSeq 2000 and HiSeq 2500 platforms.

Bioinformatics. *Sequence alignment.* Because the TLA protocol leads to reshuffling of genomic sequences, reads were mapped using our custom TLA analysis pipeline, which is based on the BWA mapping software[39]. It is a two-step process that ensures maximum 'mappability' of TLA data. First, the reads are mapped to the genome similarly to regular sequencing data. Subsequently, unaligned sequences are digested *in silico* on the basis of the NllaIII restriction site and remapped to the genome. The resulting BAM files are a combination of the first and second mapping iteration. Although we perform paired-end sequencing, we do not use the paired-end information in the mapping owing to reshuffling of the sequence. Paired ends are therefore treated separately in the general analysis.

Sampling TLA data. To assess the required sequencing depth for a successful TLA experiment, we performed random sampling of the mapped TLA data. Using SAMtools[39] we subsampled the BAM files (using the 'samtools view –s' command) and used these data to recalculated the coverage scores. By combining multiple viewpoints we can get a more even coverage distribution over the locus.

SNV calling. Because we are working with PCR amplicons we have developed a custom SNV calling pipeline. From the BAM files we create pileup files ('bwa mpileup') for our desired region. We select nucleotides that are sequenced 50 times or more. For these nucleotides we determine whether there is >25% deviation from the reference nucleotide, meaning we allow a maximum allelic imbalance of 25% (i.e., the range of diploid SNVs is between 25% and 75%, and an N_1/N_2 ratio of 0.2:0.8 would be rejected as a SNV in a diploid sample). In addition to the constraint on the allelic imbalance, we require that at least 10% of reads be in the opposite orientation. Finally, we reject SNVs that are found in the majority at the end of a sequencing read.

Titration experiment analysis. For the titration experiments we have created TLA templates in which two cell lines (SUM149PT and MDA-MB-436) are combined at various proportions (i.e., 5%/95%; 25%/75%; 50%/50%; 75%/25% and 95%/5%). We determined the differential SNVs between the two cell lines on the basis of the samples containing 95% of one cell type, using our standard SNV calling pipeline (see above). We then used this SNV list to calculate the ratio of nucleotides in the separate titration experiments. Thus, differently from in our routine SNV calling strategy, SNVs are not called *de novo* here but on the basis of an a priori list of SNPs.

De novo *haplotyping.* The basis of the *de novo* haplotyping is the idea that, for proximal sequences, the number of captures from the same allele vastly outnumbers the number from the homologous allele. For diploid samples we first perform standard SNV calling. On the basis of this list of SNVs, we can go back to the sequencing data and map the identified SNVs to the read pairs (**Fig. 1e**). Read pairs typically contain multiple NlaIII fragments, which enables linking of SNVs found in the same read pair. For a given pair of SNV positions S_i and S_j, there are a priori four possible combinations of SNVs ($S_{i,1}$ with $S_{j,1}$; $S_{i,1}$ with $S_{j,2}$; $S_{i,1}$ with $S_{j,2}$; $S_{i,2}$ with $S_{j,2}$), and two of these are the actual combination in the haplotype. To determine whether two SNVs can be significantly ($P < 0.01$) linked, we perform a Fisher exact test under the null hypothesis that there is no linkage, in which an equal distribution of the four combinations is expected. Significantly linked SNVs are then used to build the haplotype. As expected, the haplotypes that we build indeed fall apart into two distinct alleles. TLA haplotyping allows the linkage of SNVs up to at least 50 kb.

Haplotyping of chromosomal breaks is performed by SNV calling of the fused region with our standard SNV caller. As most chromosomal breaks are monoallelic, SNV calling will result in haploid SNVs, which enables the identification of the rearranged allele.

HIV sequence alignment and integration site mapping. To perform unbiased mapping of TLA data to the HIV genome, we downloaded the HIV genome compendium (http://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/compendium.html) and mapped the data to the entire compendium. We found that the subtype B.FR.83.HXB2_LAI_IIIB_BRU.K03455 was the most highly covered subtype out of the compendium and reasoned that this was the genome sequence closest to the HIV sequence present in our samples. We remapped the TLA data to this subtype to gain optimal coverage.

To map the integration sites we selected the reads from the TLA data set that contained the most flanking sequence of the long terminal repeat (LTR). After trimming the LTR sequence we remapped the remaining sequences to the genome. The unique positions were isolated, and these were selected as our set of integration sites.

34. de Vree, P.J. *et al.* Application of molecular cytogenetic techniques to clarify apparently balanced complex chromosomal rearrangements in two patients with an abnormal phenotype: case report. *Mol. Cytogenet.* **2**, 15 (2009).

doi:10.1038/nbt.2959

35. Nagel, J.H. *et al.* Gene expression profiling assigns CHEK2 1100delC breast cancers to the luminal intrinsic subtypes. *Breast Cancer Res. Treat.* **132**, 439–448 (2012).
36. Anonymous. Revision of the standards for the assessment of hormone receptors in human breast cancer; report of the second E.O.R.T.C. Workshop, held on 16–17 March, 1979, in the Netherlands Cancer Institute. *Eur. J. Cancer* **16**, 1513–1515 (1980).
37. Splinter, E., de Wit, E., van de Werken, H.J., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation. *Methods* **58**, 221–230 (2012).
38. Untergasser, A. *et al.* Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
39. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).