Targeted isolation of cloned genomic regions by recombineering for haplotype phasing and isogenic targeting

Marta Nedelkova, Marcello Maresca, Jun Fu, Maria Rostovskaya, Ramu Chenna, Christian Thiede, Konstantinos Anastassiadis, Mihail Sarov and A. Francis Stewart

1Technische Universitaet Dresden, Genomics, BiInnovationsZentrum, 2Center for Regenerative Therapies, Technische Universitaet Dresden, 3Applied Bioinformatics Group, BiInnovations Zentrum, Am Tatzaerberg 47, 4Medizinische Klinik und Poliklinik I, Universitaetsklinikum Carl Gustav Carus der Technischen Universitaet, Fetscherstrasse 74 and 5Max Planck Institute for Cell Biology and Genetics, Pfothenauerstrasse 108, D-01307 Dresden, Germany

Received February 11, 2011; Revised July 21, 2011; Accepted July 29, 2011

ABSTRACT

Studying genetic variations in the human genome is important for understanding phenotypes and complex traits, including rare personal variations and their associations with disease. The interpretation of polymorphisms requires reliable methods to isolate natural genetic variations, including combinations of variations, in a format suitable for downstream analysis. Here, we describe a strategy for targeted isolation of large regions (~35 kb) from human genomes that is also applicable to any genome of interest. The method relies on recombineering to fish out target fosmid clones from pools and thereby circumvents the laborious need to plate and screen thousands of individual clones. To optimize the method, a new highly recombineering-efficient bacterial host, including inducible TrfA for fosmid copy number amplification, was developed. Various regions were isolated from human embryonic stem cell lines and a personal genome, including highly repetitive and duplicated ones. The maternal and paternal alleles at the MECP2/IRAK 1 loci were distinguished based on identification of novel allele-specific single-nucleotide polymorphisms in regulatory regions. Additionally, we applied further recombineering to construct isogenic targeting vectors for patient-specific applications. These methods will facilitate work to understand the linkage between personal variations and disease propensity, as well as possibilities for personal genome surgery.

INTRODUCTION

Recent progress in single-nucleotide polymorphism (SNP) mapping, genome-wide association studies and massively parallel sequencing is revealing the diversity of genetic variation within the human genome (1–5). They encompass SNPs, insertions, deletions, inversions and duplications, which can be linked with disease (1,6). Understanding the genetic architecture of complex traits requires knowledge about the polymorphisms in different parts from the genome, including non-coding regions (6,7) as well as information about the haplotype phasing, that is the combination of polymorphisms at the maternal and paternal alleles (8). SNPs in intergenic and intronic elements like enhancers have been shown to regulate gene expression (9,10) and to contribute to human disorders (7,11). Recently, it was demonstrated that the activity of long interspersed elements contributes to interindividual genetic variations and can be associated with disease phenotypes (12,13).

Various methods exist for genome-wide identification of SNPs and structural variations (1). Recent advances in high-throughput DNA sequencing technologies have enabled rapid progress in the field (14) and in the near future their detection in personal genomes will be performed routinely (15,16). However, the variations lying in duplicated and highly identical sequences are still difficult to resolve and extensive bioinformatic analysis is needed to map the short next-generation sequencing reads in such regions (17,18).

Although the detection of structural variations is very important, base pair resolution of their breakpoints and further functional analysis is usually required to define their potential impact (19,20). The existing...
target-enrichment strategies, based on polymerase chain reaction (PCR) (21), hybridization or molecular inversion probes (15) merely detect variations, without isolation of the intact allele as a clone that can be further analyzed to link polymorphisms over large regions or to be genetically manipulated for downstream functional analysis. Allele linkages can be achieved using whole genome bacterial artificial chromosome (BAC) or fosmid DNA clone libraries (12,22) but the costs and time required to generate and map them are often not justified when only a specific region of the genome needs to be investigated.

In this study, we present a simple approach, based on recombineering (23,24) for targeted isolation of genomic regions in a vector format, suitable for downstream analysis. Recombineering is a DNA engineering technology, based on homologous recombination in *Escherichia coli*, mediated by the λ phage proteins Redβ/Redβ or their functional counterparts RecE/RecT from the Rac prophage (23,25). We and others have shown that recombineering has many applications, including subcloning by gap repair (25), point mutagenesis in BACs (24), oligonucleotide directed mutagenesis (26), BAC engineering for gene targeting (27,28) or protein tagging (29–31). The high efficiency and fidelity of recombineering permits high-throughput DNA engineering at genome scale (30,31).

Here, we demonstrate an application of recombineering for selective isolation of large genomic fragments of choice from complex genomes. It circumvents the need for the classical method of library screening using hybridization to filters or individually picking and end-sequencing tens of thousands of clones for indexing. The method is applicable to duplicated and repetitive regions and allows for breakpoint resolution of structural variations at single nucleotide level. The approach further allows the generation of isogenic targeting constructs with homology arms carrying the combination of SNPs characteristic for the source genome. Such constructs will facilitate genome engineering in embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs) for disease studies. We demonstrate the utility of the approach through isolation of several loci from H7 and Shef4 hES cell lines and from a cancerous genome and their subsequent haplotype variation characterization.

**MATERIALS AND METHODS**

*Escherichia coli* strains

All the strains used in this study are derived from *E. coli* DH10B. The strains GB05, GB05Red and DY380 as well as the low copy, temperature-sensitive pSC101γ/βZA plasmid were described previously (32–34). The pSC101β plasmid is derivative of pSC101γ/βZA plasmid and encodes the Redβ protein instead of the Redβγ/βRecA operon. The *E. coli* strain GB05RedTrfA was constructed by insertion of the double operon P_BADTrfA-P_RhaRedγ/βRecA at the ybeC locus of GB05 (33). For development of the cassette the P_Rha promoter was amplified from pRedFlp (30). The P_BAD promoter from the P_BADRedγ/βRecA operon was replaced with P_Rha by recombineering.

The P_BADTrfA was amplified from the genome of *E. coli* EPI300 (Epitect Biotechnologies, Madison, WI, USA) and added by recombineering to the P_RhaRedγ/βRecA.

**Stability test**

For the stability test a minimal BAC clone containing two 558 bp direct repeats was constructed from pBeloBAC11 vector [New England Biolabs (NEB), Boston, MA, USA]. The repeats are part of the chloramphenicol resistance gene (*cat*), which is split into two and is not functional. The minimal BAC clone contains also neomycin/kanamycin (*neo*) and zeocin (*zeo*) genes conferring antibiotic resistance. For the stability assay the strains were grown overnight at 30°C in LB supplemented with kanamycin (km) 10 μg/ml. From the overnight culture, 10⁶ cells were inoculated in 1 ml LB containing zeo 25 μg/ml and grown ON at 30 or 37°C. To estimate the number of spontaneous recombinants, the cells were plated on LB + chloramphenicol (cm) (15 μg/ml) and LB + zeo (15 μg/ml).

**DNA isolation and shearing**

The H7 hES DNA was prepared from cells grown in our laboratory under standard conditions. The primary bone marrow sample PS-37027 is from an acute myeloid leukemia (AML) patient. DNA was isolated applying cell lysis treatment followed by phenol–chloroform extraction, isopropanol precipitation and ethanol washing. The Shef4 hES DNA was kindly provided by Andrew Smith. The DNA was sheared using the HydroShear device (Digilab Genomic Solutions, MA, USA) and shearing assembly 4–40 kb (Zinsser Analytic, Frankfurt/Main, Germany) following the protocol for preparation of fosmid libraries (35). The sheared DNA was end-repaired and ethanol precipitated according to the metagenomic DNA isolation protocol (Epitect Biotechnologies). For library ligations between 0.4 and 1.8 kb end-repaired and precipitated DNA was used (Supplementary Table S1). The titer of the library was determined and on average 3500 clones were plated per 15-cm culture dish containing LB agar + cm (10 μg/ml) and tetracycline (tet, 5 μg/ml). Plates were incubated at 30°C for 18–24 h. To generate the pools, colonies from each dish were washed off with 2 ml LB + cm + tet, glycerol was added to 20% and 100 μl aliquots were stored at −80°C in 96-well plates. For DNA isolation from the pools, 25 μl aliquots were inoculated in 1 ml LB + cm at 37°C. The fosmids were induced to high copy overnight with 0.2% L(+)-arabinose and DNA was isolated using 96-well filter plate A (VWR International, Darmstadt, Germany). The DNA was combined in pools from one row or one column of a 96-well plate for the PCR test.
PCR pre-screen of the library

The PCR primers for pre-screening of the library (Supplementary Table S2) were designed using the Primer3 tool (http://frodo.wi.mit.edu/primer3/). The oligos were chosen to be in close proximity to the site of cassette insertion. Their sensitivity was tested with Ensembl BlastN search tool with search-sensitivity of near-exact matches and in silico PCR (http://genome.ucsc.edu/cgi-bin/hgPcr). For the PCR template, 50–100 ng DNA from each plate row or plate column was used. PCR amplification was performed using Eppendorf Mastercycler CP 534X. Thermal cycling parameters for Taq DNA polymerase (5 prime, Hamburg, Germany) were 95°C for 4 min followed by 35 cycles of 95°C for 15 s, annealing for 15 s (temperature indicated in Supplementary Table S2) and extension at 68°C for 15 s with a final extension of 10 min at 68°C. All the oligos in this study were purchased from Biomers (http://www.biomers.net/de.html).

Recombinogenic cassette design and modification

Homology arm (HA) for the capturing cassettes (Supplementary Table S3) were designed according to Ensembl (http://www.ensembl.org/index.html) genome version CRCh37 release 54–58. The cassettes were generated by PCR using the blasticidin resistance gene (bsd) and oligonucleotides that contain the flanking 50 bp homology regions. The bsd selectable marker was amplified from the genomic ara-leu locus of strain GB05 (previously recombined with this cassette) to prevent background recombination. The cassettes were phosphorylated at one 5'-end but not to the other 5'-end to generate PO or OP cassettes, where O means hydroxyl (36). The cassettes were purified from the PCR reaction using MSB Spin PCRapace kit (Invitek, Berlin, Germany). The cassette for testing the recombineering efficiency of the E. coli strains was also phosphorylated at one of the 5'-end. In addition two phosphorothioate linkages (S) were inserted in the first and second bond at the other 5'-end (PS cassette) (36).

Recombineering protocol

To screen the library by recombineering, aliquots (25 μl) from the PCR positive pools were grown in 1 ml LB supplemented with tet (5 μg/ml) and cm (10 μg/ml) overnight at 30°C. The overnight culture was diluted 1/50 and grown in 25 ml at 30°C for 2 h, followed by addition of L(+)-arabinose (Sigma A-3256) and L(+)-rhamnose (Sigma R3875) to 0.2% and growth for 45 min at 37°C. The cells were centrifuged, transferred to an Eppendorf tube and washed twice with 1 ml of ice-cold 10% glycerol, followed by resuspension in 80 μl. About 600 ng cassette was added to 40 μl competent cells. For each electroporation, a pre-chilled 1 mm electroporation cuvette (BTX, Harvard apparatus) was used at settings 1350 V, 10 μF, 600 Ω (Eppendorf Electroporator 2510). After electroporation the cells were resuspended in 1 ml SOC medium and incubated for 1 h at 37°C before plating on low-salt LB agar supplemented with 40 μg/ml blasticidin S (BSD) (InvivoGen, San Diego, CA, USA). The plates were incubated at 37°C for 18–24 h.

Characterization of the isolated recombinant fosmids

Between 1 and 16 clones per captured region were inoculated in 1 ml low salt LB supplemented with BSD 40 μg/ml and grown overnight at 37°C then 30 μl were inoculated in 0.5 ml TB supplemented with BSD and grown overnight at 37°C. To the rest, glycerol was added to 20% and stored at −80°C. Fosmid DNA was isolated by using Invirorb spin plasmid mini two (Invitex, Berlin, Germany) or 96-well filter plate A (VWR International). The clones were end-sequenced with pCC2Fos vector primers. Around 0.7 μg DNA was used for the restriction digestion experiments in a 40-μl reaction volume. All enzymes were supplied by NEB.

Next-generation sequencing parameters and bioinformatic analysis

Fosmid DNA was mixed in five pools at final concentration of ~3.5 μg/6 μl so that overlapping clones were kept in different pools. The DNA was sheared using the Covaris S2 (Covaris, Inc. Massachusetts, MA, USA) to an average fragment size of 200 bp. The fragmented pools of DNA were indexed and a standard multiplex sequencing library for Illumina platform was prepared (NEB, NEBNext® DNA Sample Preparation). After flow cell generation on the cBOT (Illumina) standard single read sequencing (51 bases) was performed on the HiSeq 2000 platform (Illumina). A total of 1.2 × 10⁶ reads were obtained from which 75% were mappable. Mapping was done with Bowtie (version 0.12.7 64-bit) against UCSC_GRCh37/hg19 human genome assembly. Initial SNP calling was carried out with samtools and subsequently custom software was written and used for the SNP analysis. The latest SNP132 database was used to annotate the variations and bambino and IGV 1.5 (Broad Institute) software was used to identify the genomic regions for polymorphisms.

Isogenic targeting constructs generation

All constructs were in silico designed using Gene Construction Kit (TEXTCO BioSoftware). The recombineering experiments were performed in the library host GB05RedTrfA, which had lost the temperature-sensitive pSC101β plasmid by culture at 37°C. The recombineering protocol was the same as described for screening the libraries but in the subsequent steps the induction was only with L(+)-rhamnose. The capturing cassettes contain 40 bp sequences flanking the bsd that serve as homology arms for sequential recombineering with the reporter cassette lacZneo (sA-T2A-LacZ-T2A-Neo-pA-loxP). The rest of the cassettes for generation of conditional knockout targeting construct were designed as already published (33). The oligos for attachment of homology arms by PCR to the capturing cassette, the sub cloning vector p15A-pTK-DTA-ampR and the downstream cassette rox-BSD-PGK-rox-loxP are given in Supplementary Table S4.
RESULTS

Generation of recombineering proficient host for fosmid library construction

Our goal was to develop an assay that can capture by recombineering large regions of interest from human genomes in a fosmid clone format suitable for sequencing and genetic engineering. We generated a new fosmid library host (GB05RedTrfA) (Figure 1), which carries in its genome the \( \gamma \beta \gamma \) RecA recombineering operon (32) under the rhamnose inducible promoter (\( P_{\text{Rha}} \)) (37) as well as the TrfA protein (38) under the arabinose inducible promoter (\( P_{\text{BAD}} \)) (39). The TrfA protein is required for initiation of the replication from the bidirectional origin OriV and subsequent increase in the fosmid copy number. The strain is highly stable (Supplementary Figure S1) with rates of spontaneous rearrangements in the absence of induction comparable with the previously published recombineering proficient hosts GB05(BAD)Red (33) or DY380 (34). We optimized the recombineering conditions using a blasticidin resistance cassette insertion assay into a single fosmid clone (Figure 1A). One of the strands of the dsDNA cassette was phosphorylated at the 5'-end and phosphothioate linkages were added to the 5'-end of the other strand, to facilitate the enzymatic conversion to ssDNA in vivo, which improves the recombineering frequencies (36).

We tested if the recombineering efficiencies can be further promoted by the helper plasmids pSC101b or pSC101\( \gamma \beta \zeta \)A (32), in which the recombineering genes are also under \( P_{\text{BAD}} \) control. The additional transient expression of the strand annealing protein Red\( \beta \) alone from the helper plasmid pSC101\( \beta \) increased the frequency of recombination almost twice as much as the additional complete recombineering operon from pSC101\( \gamma \beta \zeta \)A (Figure 1B), indicating that overexpression of some of the other proteins in the operon may be detrimental to the overall efficiency.

More than 3-fold increase in the number of recombinants was observed after high copy fosmid induction in GB05RedTrfA in comparison with the GB05Red strain where oriV cannot be induced (Figure 1B). Using the

---

Figure 1. Fosmid library host optimization. (A) Recombineering assay with GB05RedTrfA + pSC101b. The strain carries in the genome the modified red operon (32) (\( \gamma \beta \gamma \)RecA) (red) under the control of the rhamnose inducible Rha promoter and has the TrfA gene (blue), which promotes high copy fosmid replication under the control of the arabinose inducible BAD promoter. The BAD promoter also drives expression of the Red\( \beta \) protein (red) located on the helper plasmid pSC101\( \beta \). A random fosmid clone was chosen for the insertion of modified blasticidin (bsd) cassette via recombineering. The cassette is flanked with 50 bp homology arms, identical to the region of choice (green). After the selection step the temperature-sensitive plasmid pSC101b is lost at 37°C. OriV—bidirectional origin of replication, Fori—unidirectional origin of replication. (B) Strain history. All strains were derived from DH10B. The strains GB05, GB05Red and DY380 as well as the pSC101\( \gamma \beta \zeta \)A plasmid were described previously (32–34). The BADTrfA and Rha\( \gamma \beta \zeta \)A cassettes in GB05RedTrfA were inserted at \( \gamma \beta \zeta \)C locus as the BAD\( \gamma \beta \zeta \)A cassette in GB05Red. (C) Comparison of the recombineering efficiencies of the host strains. To test recombineering efficiency, a modified blasticidin cassette was inserted in a randomly selected fosmid from the H7 library. The number of recombinants was normalized to the number of cells surviving electroporation. Higher recombineering frequencies were obtained using the pSC101\( \beta \) helper plasmid. High copy fosmid induction further promoted recombineering efficiency.
GB05RedTrfA+pSC101β and transient high copy fosmid replication induction, we achieved up to $6.8 \times 10^3$ recombinants per million viable cells after transformation, an efficiency which allows for recombineering mediated targeting of a specific clone in a complex fosmid library.

**Targeted isolation of genomic regions by recombineering**

The general outline of our approach is shown in the flowchart of Figure 2. First, a fosmid library is constructed from mechanically sheared genomic DNA (Figure 2A). Next, the library is split into pools of about 3500 clones, which are then screened by PCR. Finally, the target clones are fished out by recombineering through the insertion of a modified blasticidin cassette flanked by 50-bp long homology arms (Figure 2B).

We optimized the method using genomic DNA isolated from H7 human embryonic stem (hES) cell line (40). Based on the recombineering efficiencies determined with single fosmids ($6.8 \times 10^3$ recombinants/10⁶ cells) and given that the number of surviving cells in a typical recombineering reaction in the absence of selection is about 10⁹ cells/ml, we estimated that the recombineering efficiency of the new host should allow us to isolate 10–100 recombinants of a specific clone in a mixture of 10⁴ clones. In a pilot experiment, a defined fosmid was added to pools of different complexities to determine that the optimal performance was achieved with pools of $3.5 \times 10^3$ fosmids (data not shown). At that complexity, a library of over 3-fold coverage of the haploid human genome can fit in a single 96-well plate, and any region of interest can be isolated within 2 days, saving time and effort involved in screening entire libraries.

**Application of the method to retrieve various regions**

We applied the approach to capture the OCT4 locus from the H7 hES cell line. After recombineering, blasticidin-resistant colonies were obtained from five PCR positive pools from two independent libraries (Supplementary Table S5). End sequencing from the vector and restriction analysis established that the captured fosmids covered the OCT4 locus and surrounding regions (Figure 3A; Supplementary Figure S2 and Supplementary Table S5).

Five further regions were retrieved from the H7 hES cells. For the adenosine kinase (AK), methyl CpG binding protein2 gene (MECP2) and paired box 6 (PAX6) transcriptional factor we isolated the genomic regions, required for isogenic targeting construct generation (Figure 3B–D). The entire MYCN and NANOG genes and their surrounding regions were also successfully captured (Figure 3E and F). NANOG has several pseudogenes and one of them, NANOG P1, arose through local duplication of the NANOG gene (41). In order to isolate the gene, 100 bp of homology sequence unique to the NANOG locus was chosen. The captured fosmid covers the whole locus, an intergenic region and part of the neighboring gene, which is also duplicated (41). Large parts of the 36 kb genomic fragment contain repeats from which 66% belong to different classes of Alu elements. Restriction analysis confirmed that the highly repetitive fosmids were not rearranged (Supplementary Figure S2 and Supplementary Table S5).

In further exercises, we used the male hES cell line Shef4 (42) and a primary leukemic sample. With the available cassettes, we isolated Shef4 MECP2, OCT4, PAX6 and GATA4 regions (Supplementary Table S5). For the leukemic sample, we focused on potential disease-related regions of chromosome 2 and isolated two independent clones for each of the regions of interest (TP53I3, ASXL2 and MYCNOS loci).

All target regions from both hES cells lines and the personal genome were captured successfully (Supplementary Table S5). As with other recombineering applications, we have not found any sequence limitation in the choice of homology arms except for the need to avoid repeats.
Hence the approach appears to be applicable to a diverse spectrum of genomic regions. No incorrect insertions were observed and the restriction digest analysis showed a very low number of rearranged clones. The number of recombinants varied for each of the targeted regions but was within the expected range (1–728 recombinants per reaction). Addition of more than 500 ng of the cassette did not increase the number of recombinants (Supplementary Figure S3).

We used single-strand DNA recombineering as it provides higher efficiency and fidelity (36). Either strand can be used, but the strand annealing to the lagging strand of the replication fork is favored by the recombineering reaction (43). In our experiments, the efficiencies between the two strands varied several fold (Supplementary Table S6), indicating that testing both strands can be beneficial for the isolation of difficult regions.

**Haplotype phasing and identification of allelic differences in H7 loci**

Regions from the H7 hES cell line for which more than one fosmid was fished out (Figure 3) were sequenced with Illumina in order to reconstruct the haplotype phase of the genomic regions. Indexed libraries, containing the overlapping clones were sequenced to a mean depth of 11 071 reads per base pair. Bioinformatic analyses indicated two positions on chromosome X with potential allelic differences that were supported with similar number of unique reads between the overlapping clones.

![Figure 3. Genomic regions of interest isolated from the H7 hES cell line. Depicted are the fished out clones (red), the capturing cassettes (black triangles) and the characteristics of the genomic region showing exons (thick yellow lines), introns (thin yellow lines), repetitive elements (all repeats) and G+C content (%GC). (A) Clones containing the OCT4 locus. (B) The AK locus with the capturing cassette inserted in front of exon 5. (C) Methyl CpG binding protein 2 locus (MECP2) and two different probes used for fishing out regions of interest. (D) The 3-end of the PAX6 locus. (E) Isolation of the NANOG locus from the highly similar NANOG P1 pseudogene (depicted in lilac). (F) Clone covering the MYCN gene and the probe used, which has 69% G+C content.

---

Table S6, indicating that testing both strands can be beneficial for the isolation of difficult regions.
These include differences at the MECP2/IRAK1 loci that are not annotated in SNP132 database. The observed allelic polymorphisms are G/A at the 3'UTR of MECP2 and C/G at the promoter region of IRAK1 located 5325-bp downstream on the same allele (Figure 4). Both SNPs are in CpG dinucleotides and are located in regulatory regions—a DNaseI hypersensitive site in the 3'UTR of MECP2 and the CpG island upstream of IRAK1 (USCS genome browser GRSh37/hg19). The SNP at the 3'UTR of MECP2 was validated by PCR and sequencing (data not shown). The second SNP is located in an extremely GC rich region and we failed to amplify it by PCR with several sets of primers.

In addition to the allele-specific SNPs we reconstructed the combination of SNPs across the sequenced regions of chromosome X, 6 and 10 (Supplementary Table S7). As expected more SNPs were found in the highly polymorphic region of chromosome 6 than at the other loci. In addition several non-synonymous mutations in CCHCR1 and TCF19 genes and small-scale indels were scattered across the 35 kb genomic region from chromosome 6 (data not shown). The indels for the OCT4 loci from the H7 and Shef4 cell line were validated by PCR and sequencing (Supplementary Table S8).

**DISCUSSION**

Studying genetic variations in the human genome is important for the understanding of phenotypes, diseases, drug responsiveness and the mechanisms of complex traits (6). For many applications, only a small part of the genome, such as specific genes or regulatory regions, are of interest (44,45). The current methods for selected enrichment of genomic regions followed by next generation sequencing are based on PCR or hybridization approaches (15). These methods encounter size limitations particularly to link variations separated by more than a few hundred base pairs, as well as limitations in duplicated and repetitive regions.

The recombineering strategy presented here is useful for targeted isolation of genomic regions in a vector format that allows for rapid adaptation to functional analysis based on gene targeting (27,28) or transgenesis (30). A similar approach to isolate genomic regions in BACs has been published recently (46). We use fosmids, because they are easy to handle, stable, suitable for genomic structural variation studies (2,5,22) and preparation of targeting constructs. Most importantly, compared to BAC libraries, fosmid library construction requires much less genomic DNA, which is a major consideration when the source of DNA is a patient sample.

![Figure 4. Allele-specific SNPs at the MECP2/IRAK1 locus on the X chromosome in H7 hES cells. In the upper part of the diagram the distribution of uniquely mappable Illumina reads from fosmid clones H7-F and H7-C02 is shown as grey lines. Gaps indicate repetitive sequence and the average corresponds to 11 000 reads. The colored lines indicate the SNP positions in comparison with the human reference genome as follows: green-A; red-T; blue-C and orange-G. The lower part of the diagram shows a 14 kb region containing two allele-specific SNPs among seven other SNPs and one newly identified indel, which is present on both alleles. One of the alleles has A instead of G at position 153 290 956 and G instead of C at position 153 285 631. The SNPs are in DNaseI hypersensitive site (DHS) and in CpG island, respectively as indicated. The indel indicates an insertion of a G onto a poly G track at 153 287 325.](https://academic.oup.com/nar/article-abstract/39/20/e137/2409837)
To increase the targeting efficiency and thereby the complexity of the pools from which a specific region can be retrieved, we engineered a new strain that allows for switching from unidirectional to bidirectional fosmid replication. In that way, we exploit an additional increase in recombineering efficiency due to increased fosmid copy-number after TrfA induction. This improved the isolation of genomic regions of choice from complex fosmid pools. The very low levels of illegitimate recombination reduced the need to screen through a large number of clones to obtain the desired region. The number of recombinants varied between the captured loci, possibly reflecting the different replication speeds of the individual clones within the pools. Variability in the number of recombinants for several E. coli chromosomal locations has previously been correlated with the rate of replication of the regions (26).

To increase the targeting efficiency and thereby the complexity of the pools from which a specific region can be retrieved, we engineered a new strain that allows for switching from unidirectional to bidirectional fosmid replication. In that way, we exploit an additional increase in recombineering efficiency due to increased fosmid copy-number after TrfA induction. This improved the isolation of genomic regions of choice from complex fosmid pools. The very low levels of illegitimate recombination reduced the need to screen through a large number of clones to obtain the desired region. The number of recombinants varied between the captured loci, possibly reflecting the different replication speeds of the individual clones within the pools. Variability in the number of recombinants for several E. coli chromosomal locations has previously been correlated with the rate of replication of the regions (26).

Previously a method to screen genomic libraries by recombineering was reported (47). However, this method does not appear to have been subsequently utilized, possibly because the complex counter selection strategy imposed practical difficulties. Similarly our previous experience with genomic cloning by recombineering (25), indicated certain practical limits to lambda Red recombination in complex backgrounds. Hence, we adapted a recombineering method to optimally sized pools of cloned genomic regions.

Fine-tuning the expression levels of the recombineering proteins not only improved the recovery of target clones but also likely contributed to the successful isolation of intact, highly repetitive, regions. We adapted a recombineering method to optimally sized pools of cloned genomic regions.

To increase the targeting efficiency and thereby the complexity of the pools from which a specific region can be retrieved, we engineered a new strain that allows for switching from unidirectional to bidirectional fosmid replication. In that way, we exploit an additional increase in recombineering efficiency due to increased fosmid copy-number after TrfA induction. This improved the isolation of genomic regions of choice from complex fosmid pools. The very low levels of illegitimate recombination reduced the need to screen through a large number of clones to obtain the desired region. The number of recombinants varied between the captured loci, possibly reflecting the different replication speeds of the individual clones within the pools. Variability in the number of recombinants for several E. coli chromosomal locations has previously been correlated with the rate of replication of the regions (26).

Previously a method to screen genomic libraries by recombineering was reported (47). However, this method does not appear to have been subsequently utilized, possibly because the complex counter selection strategy imposed practical difficulties. Similarly our previous experience with genomic cloning by recombineering (25), indicated certain practical limits to lambda Red recombination in complex backgrounds. Hence, we adapted a recombineering method to optimally sized pools of cloned genomic regions.

To increase the targeting efficiency and thereby the complexity of the pools from which a specific region can be retrieved, we engineered a new strain that allows for switching from unidirectional to bidirectional fosmid replication. In that way, we exploit an additional increase in recombineering efficiency due to increased fosmid copy-number after TrfA induction. This improved the isolation of genomic regions of choice from complex fosmid pools. The very low levels of illegitimate recombination reduced the need to screen through a large number of clones to obtain the desired region. The number of recombinants varied between the captured loci, possibly reflecting the different replication speeds of the individual clones within the pools. Variability in the number of recombinants for several E. coli chromosomal locations has previously been correlated with the rate of replication of the regions (26).

Previously a method to screen genomic libraries by recombineering was reported (47). However, this method does not appear to have been subsequently utilized, possibly because the complex counter selection strategy imposed practical difficulties. Similarly our previous experience with genomic cloning by recombineering (25), indicated certain practical limits to lambda Red recombination in complex backgrounds. Hence, we adapted a recombineering method to optimally sized pools of cloned genomic regions.

To increase the targeting efficiency and thereby the complexity of the pools from which a specific region can be retrieved, we engineered a new strain that allows for switching from unidirectional to bidirectional fosmid replication. In that way, we exploit an additional increase in recombineering efficiency due to increased fosmid copy-number after TrfA induction. This improved the isolation of genomic regions of choice from complex fosmid pools. The very low levels of illegitimate recombination reduced the need to screen through a large number of clones to obtain the desired region. The number of recombinants varied between the captured loci, possibly reflecting the different replication speeds of the individual clones within the pools. Variability in the number of recombinants for several E. coli chromosomal locations has previously been correlated with the rate of replication of the regions (26).

Previously a method to screen genomic libraries by recombineering was reported (47). However, this method does not appear to have been subsequently utilized, possibly because the complex counter selection strategy imposed practical difficulties. Similarly our previous experience with genomic cloning by recombineering (25), indicated certain practical limits to lambda Red recombination in complex backgrounds. Hence, we adapted a recombineering method to optimally sized pools of cloned genomic regions.

To increase the targeting efficiency and thereby the complexity of the pools from which a specific region can be retrieved, we engineered a new strain that allows for switching from unidirectional to bidirectional fosmid replication. In that way, we exploit an additional increase in recombineering efficiency due to increased fosmid copy-number after TrfA induction. This improved the isolation of genomic regions of choice from complex fosmid pools. The very low levels of illegitimate recombination reduced the need to screen through a large number of clones to obtain the desired region. The number of recombinants varied between the captured loci, possibly reflecting the different replication speeds of the individual clones within the pools. Variability in the number of recombinants for several E. coli chromosomal locations has previously been correlated with the rate of replication of the regions (26).

Previously a method to screen genomic libraries by recombineering was reported (47). However, this method does not appear to have been subsequently utilized, possibly because the complex counter selection strategy imposed practical difficulties. Similarly our previous experience with genomic cloning by recombineering (25), indicated certain practical limits to lambda Red recombination in complex backgrounds. Hence, we adapted a recombineering method to optimally sized pools of cloned genomic regions.

To increase the targeting efficiency and thereby the complexity of the pools from which a specific region can be retrieved, we engineered a new strain that allows for switching from unidirectional to bidirectional fosmid replication. In that way, we exploit an additional increase in recombineering efficiency due to increased fosmid copy-number after TrfA induction. This improved the isolation of genomic regions of choice from complex fosmid pools. The very low levels of illegitimate recombination reduced the need to screen through a large number of clones to obtain the desired region. The number of recombinants varied between the captured loci, possibly reflecting the different replication speeds of the individual clones within the pools. Variability in the number of recombinants for several E. coli chromosomal locations has previously been correlated with the rate of replication of the regions (26).

Previously a method to screen genomic libraries by recombineering was reported (47). However, this method does not appear to have been subsequently utilized, possibly because the complex counter selection strategy imposed practical difficulties. Similarly our previous experience with genomic cloning by recombineering (25), indicated certain practical limits to lambda Red recombination in complex backgrounds. Hence, we adapted a recombineering method to optimally sized pools of cloned genomic regions.

Figure 5. Workflow for the generation of isogenic conditional targeting constructs. All recombineering steps after the first one were performed in GB05RedTrfA after rhamnose induction of the Red/βzA operon. All the genes conveying antibiotic resistance have prokaryotic promoters (data not shown). The antibodies used for selection of the recombinants at each stage are indicated at the right of each step: cm, chloramphenicol; bsd, blasticidin; km, kanamycin; amp, ampicillin. (A) The fosmid clone was identified by insertion of the bsd (lilac) as described above using 50 bp homology arms to the genomic region of interest (orange) and thereby inserting the homology arms for the next recombineering step (blue). The insertion site was chosen to be nearby a frame-shifting exon (orange). Fori and oriV, fosmid replication origins; cat, chloramphenicol resistance gene. (B) The bsd gene is replaced by a lacZneo reporter/stop cassette that conveys kanamycin resistance by recombination through the homology arms (blue). The lacZneo cassette contains our standard design (27,28,33) including frt sites for FLP recombinase and a loxp site for Cre recombinase at the 3’-end. The cassette also contains splice acceptor (sA); ribosomal-skipping peptides (t2A) and the SV40 polyadenylation signal (SVpA). Neo is the kanamycin/G418 resistance gene and lacZ is the β-galactosidase gene. (C) A region containing the lacZneo cassette, the frame shifting exon and flanking ~2 kb homology arms is subcloned by gap repair into a p15A plasmid vector. The plasmid vector also includes the HSVtk promoter (TK) driven diphtheria toxin A chain gene (DTA) for counter selection and the ampicillin resistance gene (bla). (D) A second loxp site is added on the 3’ side of the frameshifting exon using blasticidin selection. The PGK–BSD gene is flanked by rox sites for later removal by Dre recombinase.
The extent of variation within human genomes is now being revealed by SNP maps and massively parallel sequencing (1–4). However, knowledge about the ‘haplotype phasing’ in different genomes has been scarce (8). Two recently published methods for genome-wide resolution of the haplotypes (49,50) pave the way to systematically study haplotype phasing in individual genomes and cell lines. Our approach is complementary to these studies and allows for the determination of SNP linkage and therefore the disease susceptibility throughout the selected regions covered by fosmid clones. Thereby, we reconstructed haplotypes at loci from chromosome 6, X and 10 from the H7 hES cell line. Comparative analysis between the H7 and Shef4 OCT4 haplotypes revealed differences in 12 SNP positions and most of the identified indels were cell line specific (13 of 16). These variations were found in more than one independent clone and therefore represent true polymorphisms of the cell lines.

Whole-genome sequencing shows that structural variations smaller than 50 kb account for the large portion of polymorphism identified in individual human genomes (1,5). Most of these events are enriched near or in repeated and segmental duplicated regions and difficulties to resolve them have been reported by different investigators (5,17). Using the targeted retrieval of clones, we were able to distinguish between highly similar sequences like NANOG and its pseudogene NANOG P1. Once isolated, such regions can be further characterized by sequencing at very high depth. This allows the description of their polymorphisms at single nucleotide resolution.

Exploring the impact of the mutations and their characterization as benign or disease associated can be achieved through gene targeting in stem cells (51,52) with isogenic constructs. Our approach permits generation of such constructs with personal genome specific combination of variations. The isogenicity of the flanking homologous sequences is an important issue. First, it could promote the targeting efficiency in human ES cells as was shown for mouse ES cells (48,53). Second, bearing in mind that SNPs may influence transcription factor binding and gene expression (9,10), targeting with isogenic vectors should not disturb the existing genomic context. This will be useful for gene editing in stem cell-based therapies.

We identified two novel allele-specific SNPs located in regulatory regions on one of the X chromosome in the H7 cell line at the MECP2/IRAK1 loci. The biological significance of these polymorphisms is not known. The whole-genome ENCODE analysis on the male H1 hES cell line indicates that the two SNPs are located in an enhancer and a promoter where c-Myc and Pol2 bind, respectively. The SNPs are in CpG dinucleotides thus they may influence the binding of regulatory proteins or the methylation status of the two alleles.

The high fidelity of Red/ET recombineering demonstrated in this and previous studies allows the further scale up of the method to high-throughput liquid format (30,31) for simultaneous isolation of multiple loci. For example, the method can be used to develop screening assays for isolation of regions affected by mobilized retrotransposons or other repetitive elements in personal genomes. Recently, numerous novel active retrotransposons were identified in the human genome (12,13). Although they are underrepresented in the reference sequence, they exist at low allele frequencies in the population and can be a source for disease-producing insertions.

This method can also simplify the acquisition of DNA regions from model organisms or metagenomic studies of environmental samples. The approach is straightforward and does not require any special equipment or complicated computational analysis. Because it is flexible with many potential applications, we recommend it to a wide range of researchers.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
The authors thank Andrew Smith for providing the DNA from Shef4 hES cell line. The authors are grateful to Andreas Dahl of the Deep Sequencing Facility at Biotec Dresden, for Illumina sequencing and the primary data analysis. M.N. designed and performed the experiments for all the main figures and most of the Supplementary material. J.F. performed the experiments for Supplementary Fig. 1. M.R. cultured the H7 hES cell line. R.C. did the bioinformatic analysis for Supplementary Table 7. C.T. provided the leukemic sample. M.S., K.A., M.M., J.F. and A.F.S. contributed with ideas and discussions throughout the project. M.N. and A.F.S. prepared the manuscript.

FUNDING
NGFN plus (Nationales Genomforschungsnetz of the Bundesministerium fuer Bildung und Forschung; 01GS0872) (to C.T. and A.F.S.); European Commission 6th Framework Program, ESTOOLS and 7th Framework Program, EUCOMMTOOLS (to A.F.S.). Funding for open access charges: NGFN program of the Bundesministerium fuer Bildung und Forschung Leukemia grant (to C.T. and A.F.S.).

Conflict of interest statement. The primary patents for recombineering are held by Gene Bridges GmbH, a company which AFS founded and is a major shareholder.

REFERENCES
1. Feuk,L., Carson,A.R. and Scherer,S.W. (2006) Structural variation in the human genome. Nat. Rev. Genet., 7, 85–97.
2. Kidd,J.M., Cooper,G.M., Donahue,W.F., Hayden,H.S., Sampas,N., Graves,T., Hansen,N., Teague,B., Alkan,C., Antonacci,F. et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature, 453, 56–64.
3. Lander,E.S. (2011) Initial impact of the sequencing of the human genome. Nature, 470, 187–197.
4. Mills,R.E., Walter,K., Stewart,C., Handsaker,R.E., Chen.K., Alkan,C., Abyzov,A., Yoon,S.C., Ye,K., Cheetham,R.K. et al.
21. Tewhey, R., Warner, J.B., Nakano, M., Libby, B., Medkova, M., Kidd, J.M., Newman, T.L., Tuzun, E., Kaul, R. and Eichler, E.E. (2009) Human genome variation and its contribution to complex traits. *Nat. Rev. Genet.*, 10, 241–251.

22. Visel, A., Rubin, E.M. and Pennacchio, L.A. (2009) Genomic views of distant-acting enhancers. *Nature*, 461, 199–205.

23. Bansal, V., Tewhey, R., Topol, E.J. and Schork, N.J. (2011) The next phase in human genetics. *Nat. Biotechnol.*, 29, 38–39.

24. Bandele, O.J., Wang, X., Campbell, M.R., Pittman, G.S. and Stewart, A.F. (2000) Point mutation of bacterial artificial chromosomes to create multipurpose alleles. *Nat. Biotechnol.*, 18, 1314–1317.

25. Ellis, H.M., Yu, D., DiTizio, T. and Court, D.L. (2001) High efficiency mutagenesis, repair, and engineering of chromosomal DNA using single-stranded oligonucleotides. *Proc. Natl Acad. Sci. USA*, 98, 6742–6746.

26. Testa, G., Zhang, Y., Vintersten, K., Benes, V., Pijnappel, W.W., Chambers, J., Smith, A.J., Smith, A.G. and Stewart, A.F. (2003) Engineering the mouse genome with bacterial artificial chromosomes to create multipurpose alleles. *Nat. Biotechnol.*, 21, 443–447.

27. Skarnes, W.C., Rosen, B., West, A.P., Koutou sirakis, M., Bushell, W., Iyer, V., Mujica, A.O., Thomas, M., Harrow, J., Cox, T. et al. (2011) A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, 474, 337–342.

28. Hofemeister, H., Ciotta, G., Fu, J., Seibert, P., Schulz, A., Maresca, M., Sarov, M., Anastassiads, K. and Stewart, A.F. (2011) Recombineering, transfection, Western and ChIP methods for protein tagging via gene targeting or BAC transgenesis. *Methods*, 53, 437–452.

29. Sarov, M., Schneider, S., Pozniakowski, A., Roguev, A., Ernst, S., Zhang, Y., Hyman, A.A. and Stewart, A.F. (2006) A recombineering pipeline for functional genomics applied to Caenorhabditis elegans. *Nat. Methods*, 3, 839–844.

30. Poser, I., Sarov, M., Hutchins, J.R., Heriche, J.K., Toyoda, Y., Pozniakowski, A., Weigl, D., Nitzsche, A., Heggemann, B., Bird, A.W. et al. (2008) BAC TransgeneOmis: a high-throughput method for exploration of protein function in mammals. *Nat. Methods*, 5, 409–415.

31. Wang, J., Sarov, M., Rientjes, J., Fu, J., Hollak, H., Kranz, H., Xie, W., Stewart, A.F. and Zhang, Y. (2006) An improved recombineering approach by adding RecA to lambda Red recombination. *Mol. Biotechnol.*, 32, 43–53.

32. Fu, J., Teucher, M., Anastassiads, K., Skarnes, W. and Stewart, A.F. (2010) A recombineering pipeline to make conditional targeting constructs. *Methods Enzymol.*, 477, 125–144.

33. Lee, E.C., Yu, D., Martinez de Velasco, J., Tessarollo, L., Swang, D.A., Court, D.L., Jenkins, N.A. and Copeland, N.G. (2001) A highly efficient *Escherichia coli*-based chromosome engineering system adapted for recombinein targetg and subcloning of BAC DNA. *Genomics*, 73, 56–65.

34. Donahue, W.F. and Ebling, H.M. (2007) Fosmid libraries for genomic structural variation detection. *Curr Protoc Hum Genet*, Chapter 5, Unit 5.20.

35. Maresca, M., Erler, A., Fu, J., Friedrich, A., Zhang, Y. and Stewart, A.F. (2010) Single-stranded heteroduplex intermediates in lambda Red homologous recombination. *BMC Mol. Biol.*, 11, 54.

36. Cardona, S.T. and Valvano, M.A. (2005) An expression vector containing a thamnos-inducible promoter provides tightly regulated gene expression in *Buchhorderia cenoeca*. *Plasmid*, 54, 219–228.

37. Wild, J., Hradecna, Z. and Zsybalski, W. (2002) Conditionally amplifiable BACs: switching from single-copy to high-copy vectors and genomic clones. *Genome Res.*, 12, 1434–1444.

38. Guzman, L.M., Belin, D., Carson, M.J. and Beckwith, J. (1995) Tight regulation, modulation and high-level expression by vectors containing the arabinoibis PBAD promoter. *J. Bacteriol.*, 177, 4121–4130.

39. Thomoson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waktizn, M.A., Swiergle, J.J., Marshall, V.S. and Jones, J.M. (1998) Embryonic stem cell lines derived from human blastocysts. *Science*, 282, 1145–1147.

40. Booth, H.A. and Holland, P.W. (2004) Eleven daughters of NANOGenomics. *Genomics*, 84, 229–238.

41. Allen, G.A., Massion, S., Haines, A., Chao, D. and Moore, H. (2010) Generation of Sheffiel (Shed) human embryonic stem cell lines using a microdrop culture system. *Plasmid*, 63, 249–259.

42. Zhang, Y., Muyers, J.P., Rientjes, J. and Stewart, A.F. (2003) Phage annealing proteins promote oligonucleotide-directed mutagenesis in *Escherichia coli* and mouse ES cells. *BMC Mol. Biol.*, 4, 1.

43. Nijman, J.J., Mokry, M., van Boxtel, R., Toonen, P., de Bruijn, E. and Cuppen, E. (2010) Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. *Nat. Methods*, 7, 913–915.
45. Chmielecki, J., Peifer, M., Jia, P., Socci, N.D., Hutchinson, K., Viale, A., Zhao, Z., Thomas, R.K. and Pao, W. (2010) Targeted next-generation sequencing of DNA regions proximal to a conserved GXGXXG signaling motif enables systematic discovery of tyrosine kinase fusions in cancer. *Nucleic Acids Res.*, 38, 6985–6996.

46. Nefedov, M., Carbone, L., Field, M., Schein, J. and de Jong, P.J. (2011) Isolation of specific clones from nonarrayed BAC libraries through homologous recombination. *J. Biomed. Biotechnol.*, 2011, 560124.

47. Zhang, P., Li, M.Z. and Elledge, S.J. (2002) Towards genetic genome projects: genomic library screening and gene-targeting vector construction in a single step. *Nat. Genet.*, 30, 31–39.

48. Yang, Y. and Seed, B. (2003) Site-specific gene targeting in mouse embryonic stem cells with intact bacterial artificial chromosomes. *Nat. Biotechnol.*, 21, 447–451.

49. Fan, H.C., Wang, J., Potanina, A. and Quake, S.R. (2011) Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.*, 29, 51–57.

50. Kitzman, J.O., Mackenzie, A.P., Adey, A., Hiatt, J.B., Patwardhan, R.P., Sudmant, P.H., Ng, S.B., Alkan, C., Qiu, R., Eichler, E.E. *et al.* (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.*, 29, 59–63.

51. Buecker, C., Chen, H.H., Polo, J.M., Daheron, L., Bu, L., Barakat, T.S., Okwieka, P., Porter, A., Gribnau, J., Hochedlinger, K. *et al.* (2010) A murine ESC-like state facilitates transgenesis and homologous recombination in human pluripotent stem cells. *Cell Stem Cell*, 6, 535–546.

52. Song, H., Chung, S.K. and Xu, Y. (2010) Modeling disease in human ESCs using an efficient BAC-based homologous recombination system. *Cell Stem Cell*, 6, 80–89.

53. Zhou, L., Rowley, D.L., Mi, Q.S., Sefcovic, N., Matthes, H.W., Kieffer, B.L. and Donovan, D.M. (2001) Murine inter-strain polymorphisms alter gene targeting frequencies at the mu opioid receptor locus in embryonic stem cells. *Mamm. Genome*, 12, 772–778.