Chapter 13

A Fosmid Pool-Based Next Generation Sequencing Approach to Haplotype-Resolve Whole Genomes

Eun-Kyung Suk, Sabrina Schulz, Birgit Mentrup, Thomas Huebsch, Jorge Duitama, and Margret R. Hoehe

Abstract

Haplotype resolution of human genomes is essential to describe and interpret genetic variation and its impact on biology and disease. Our approach to haplotyping relies on converting genomic DNA into a fosmid library, which represents the entire diploid genome as a collection of haploid DNA clones of ~40 kb in size. These can be partitioned into pools such that the probability that the same pool contains both parental haplotypes is reduced to ~1%. This is the key principle of this method, allowing entire pools of fosmids to be massively parallel sequenced, yielding haploid sequence output. Here, we present a detailed protocol for fosmid pool-based next generation sequencing to haplotype-resolve whole genomes including the following steps: (1) generation of high molecular weight DNA fragments of ~40 kb in size from genomic DNA; (2) fosmid cloning and partitioning into 96-well plates; (3) barcoded sequencing library preparation from fosmid pools for next generation sequencing; and (4) computational analysis of fosmid sequences and assembly into contiguous haploid sequences.

This method can be used in combination with, but also without, whole genome shotgun sequencing to extensively resolve heterozygous SNPs and structural variants within genomic regions, resulting in haploid contigs of several hundred kb up to several Mb. This method has a broad range of applications including population and ancestry genetics, the clinical interpretation of mutations in personal genomes, the analysis of cancer genomes and highly complex disease gene regions such as MHC. Moreover, haplotype-resolved genome sequencing allows description and interpretation of the diploid nature of genome biology, for example through the analysis of haploid gene forms and allele-specific phenomena. Application of this method has enabled the production of most of the molecular haplotype-resolved genomes reported to date.

Key words Haplotype-resolving genomes, Molecular haplotypes, Phasing, Clone-based haplotyping, Fosmid library, Fosmid pools, Fosmids, Next generation sequencing, Haplotype assembly, Phasing algorithm

1 Introduction

Human genomes are diploid by nature. Thus, to fully understand human biology and link genetic variation to gene function and phenotype, it is essential to determine both parental sequences of an individual genome independently [1, 2]. Present technologies,

Irene Tiemann-Boege and Andrea Betancourt (eds.), *Haplotyping: Methods and Protocols*, Methods in Molecular Biology, vol. 1551, DOI 10.1007/978-1-4939-6750-6_13, © Springer Science+Business Media LLC 2017

however, routinely read out "mixed diploid" sequences. Therefore, they cannot distinguish between the unique combinations of variants on each of the two chromosomal homologues, the haplotypes. Over the past few years, a number of experimental methods to haplotype-resolve genomes have been developed. Among those, clone-based haplotyping, in particular fosmid pool-based next generation sequencing (NGS), has enabled phasing by far the largest number of genomes to date, over 30 [3–9]. Here, we present the principle and concrete steps of this method.

The key principle is to convert human genomic DNA into a library of fosmids, haploid DNA fragments ~40 kilobases (kb) in size, and partition this library into pools of fosmids such that the probability that both parental haplotypes co-occur is reduced to ~1% [10]. Thus, multiple pools can be massively parallel sequenced to generate redundant coverage of both haploid genomes of an individual. In our original report introducing this principle [10], we have established "haploid clone pools" of ~5000 fosmids, random mixtures of DNA fragments representing ~5% of a haploid genome. These result from partitioning a library of $\sim 1.44 \times 10^6$ fosmid clones, ensuring 7× coverage of each haploid genome, into 3×96-well plates. In order to increase the throughput, we have chosen to combine these plates into one 96-well plate, each well containing a super-pool of ~15,000 fosmids. These super-pools are barcoded and subjected to NGS. Analysis of the NGS data showed that only 1.31% of SNP calls, on average, were heterozygous per super-pool, confirming that the method works as expected [3]. To estimate the number of super-pools necessary to be sequenced in order to reach a sufficient coverage level, simulation studies were performed. Accordingly, 40 pools were estimated to result in an average haploid read coverage of ~12× (diploid read coverage of ~24×) and 85% of heterozygous SNPs phased; 48 pools were required to achieve a read coverage of 14.5× and ~29×, respectively, and 92% of SNPs phased.

In the molecular genetics part of our protocol, we describe the following steps (overview in Fig. 1): (1) Extraction of high molecular weight (HMW) genomic DNA (gDNA), mechanical shearing, and gel-based selection of DNA fragments of ~40 kb in size; (2) ligation of size-selected fragments into pEpiFos vector; (3) phage packaging and mass transfection of *E. Coli* to obtain a total of 1.44×10^6 fosmid clones; (4) partitioning of these fosmid clones into 3×96 deep well plates to generate pools of ~5000 fosmids and amplifying those in liquid culture; (5) combining the 3×96 well plates to generate super-pools of 15,000 fosmid clones; (6) amplification of fosmid clones per well on agar plates; (7) isolation of fosmid DNA from amplified clones; (8) preparation of barcoded sequencing libraries per super-pool, and (9) processing them for NGS. We also indicate where the protocol can be adapted to newer sequencing platforms, so that it is clear which steps to modify. This

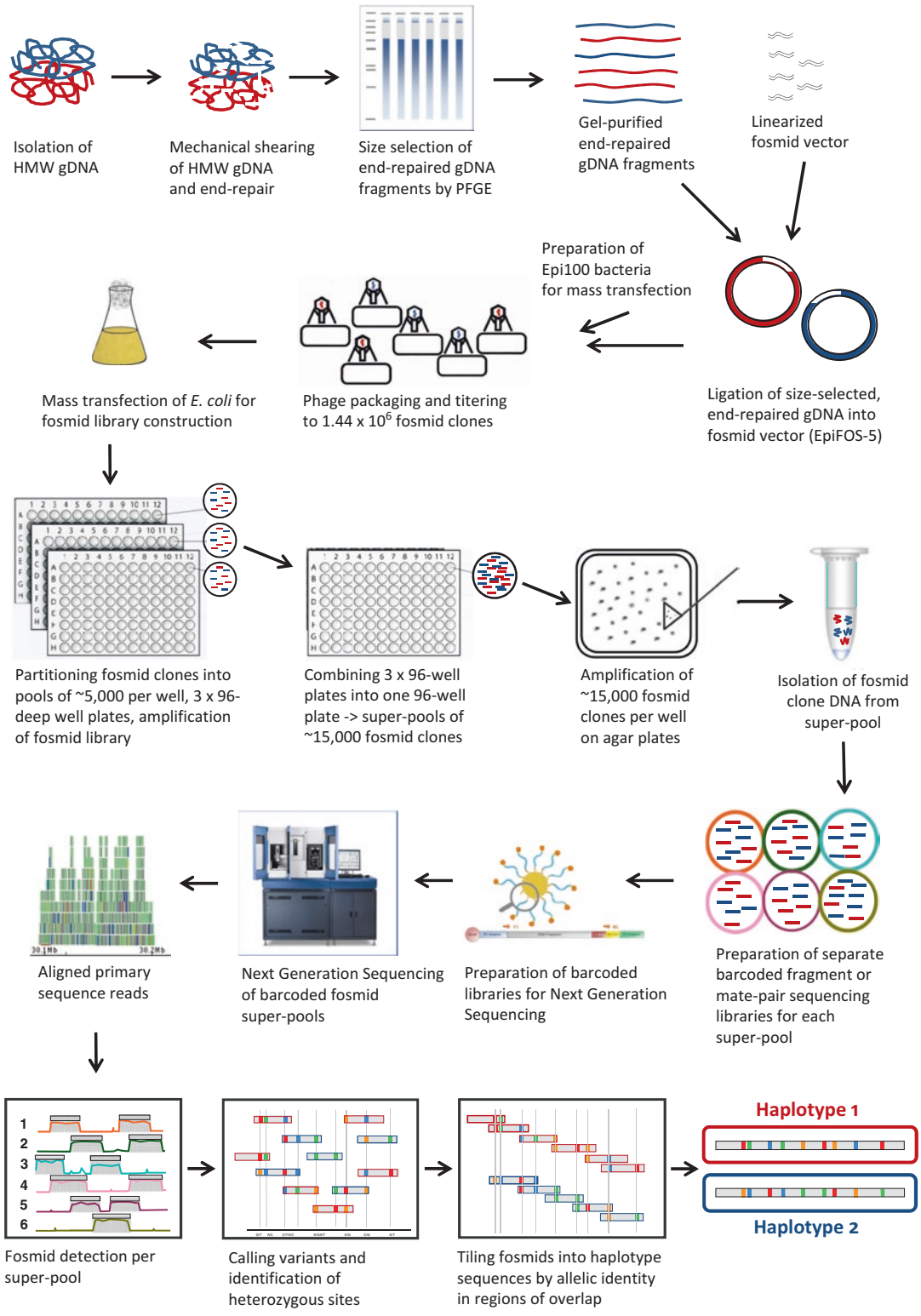

Fig. 1 Overview of fosmid pool-based Next Generation Sequencing method to haplotype-resolve whole genomes. The principle and molecular genetics and computational key steps of this method, as applied to haplotype-resolving one individual human genome, are shown

protocol can be complemented by any routine protocol for whole genome shotgun sequencing (WGS), as we have done to generate our first haplotype-resolved genome, "Max Planck One" (MP1) [3]. This will add to the quality of heterozygote detection and help to resolve larger structural variants. Successive production of a set of 12 genomes has shown that assembly of haplotype contigs solely from fosmid pool-based NGS is feasible [9].

In the computational part of our protocol, we describe the steps involved in assembling a haplotype-resolved genome from the sequence read output in a user-oriented mode, providing downloadable algorithms and scripts. In particular, we make our specifically designed heuristic algorithm "ReFHap" [11] available, which has proven to be particularly efficient in terms of computing time and represents an optimal compromise between accuracy, completeness, and computational resources [5]. Our computational tools should be usable by any scientist with experience in NGS data analysis. Haplotype assembly includes the following steps (*see* Fig. 6 in the bioinformatics part): (1) alignment of fosmid pool sequence reads and sorting these alignments per pool; (2) merging sorted alignments to detect phase-informative, heterozygous variants; (3) fosmid detection and allele calling per pool; and (4) phasing by the use of ReFHap to generate haplotype contigs of phased variants per chromosome.

Our method was empirically corroborated by application to haplotyping HapMap trio child NA12878 [5]. For this sample, whole genome sequencing of the family trio had resulted in the resolution of ~80% of all heterozygous positions [12]. Where comparable, the trio-based and molecular phasing data were entirely identical, showing that fosmid pool-based haplotyping can produce highly accurate results, even at low coverage. Our method, however, allowed resolution of a much higher number of heterozygous SNPs, ~98% in total, highlighting the power of our approach. Moreover, our method enabled generation of the most comprehensively haplotype-resolved human genome to date, MP1, with over 99% of all heterozygous SNPs and virtually all individual and rare SNPs phased into long haplotype blocks with an N50 of ~1 Megabase (Mb), i.e., 50% of the haplotype-resolved sequence were within blocks of at least ~1 Mb. The maximum block length achieved was ~6.3 Mb [3]. Finally, we have applied our protocol to phase an entire set of 16 human genomes [9], equivalent to about half of the published production of clone pool-based haplotype-resolved genomes.

In sum, haplotype-resolved genome sequencing is applicable to a broad range of scenarios, including population and anthropological genetics and the analysis of human diseases, for example the determination of the molecular haplotypes underlying highly variable and complex disease gene regions such as MHC, or GWAS regions. Moreover, knowledge of phase is critical for clinical

interpretation of mutations such as compound heterozygosity or pharmacogenetically relevant variants. The same is true for the accurate and actionable description and interpretation of personal genomes as a whole. Finally, haplotype-resolved genome sequencing provides key information on genome function, for example through resolving haploid gene forms and their regulatory environments, and allele-specific phenomena in general. These include gene expression, the regulation of transcription, methylation, and other epigenetic mechanisms.

2 Materials

2.1 Fosmid Library Construction

2.1.1 High Molecular Weight (HMW) Genomic DNA (gDNA) Isolation

1. DNA Extraction Kit for HMW gDNA (Stratagene, Cat. Nr. 200600).
2. Centrifuge (12,500 × *g*) for 15 ml/50 ml tubes.
3. Centrifuge (refrigerated, 14,000 × *g*) for 1.5/2.0 ml tubes and 96-well plates.
4. Shaking Water bath.
5. Nanodrop ND100.
6. Rotating tube roller.
7. 50 ml reaction tubes.
8. 1.5/2 ml reaction tubes.
9. Pipettors and large-bore pipet tips.
10. Sterile glass rod.
11. Absolute EtOH p.A.
12. 70% EtOH.
13. TE buffer: 10 mM Tris–HCl (pH 7.5), 1 mM EDTA.
14. Pulsed Field Gel Electrophoresis System (CHEF DR II).
15. Ultrapure Low Melting Point Agarose (Thermo Fisher).
16. 10× TBE: [890 mM Tris, 890 mM Boric acid, 20 mM EDTA (pH 8.0)]; 0.5× TBE.
17. Lambda DNA-Mono Cut Mix DNA Ladder (NEB).
18. Loading dye.
19. Ethidium bromide.
20. UV-illuminator.

2.1.2 Mechanical Shearing of HMW gDNA and End-Repair

1. HydroShear shearing device (Digilab Genomic Solutions).
2. Sterile water.
3. Sterile disposable 1 ml syringes, 23G × 1¼ injection needle.
4. Agarose gel electrophoresis system.
5. 100% Isopropanol p.A.

6. TE buffer: 10 mM Tris–HCl (pH 7.5), 1 mM EDTA.
7. EpiFOS Fosmid Library Production Kit (Epicentre, Cat. No. FOS0901).
8. 1.5 ml reaction tubes.
9. Thermomixer.

2.1.3 Size Selection of End-Repaired gDNA Fragments by PFGE

1. Pulsed Field Gel Electrophoresis System (CHEF DR II).
2. 1% Low melting agarose.
3. 0.5× TBE.
4. 6× Loading Dye.
5. Lambda DNA-Mono Cut Mix DNA Ladder (NEB).
6. Sterile water.
7. Gel chamber (14 cm wide, 13 cm long) with preparative comb (4 wells; 1.5 mm thick, 27 mm width, 14 mm length, 200 µl volume/well).
8. SYBR® Gold Nucleic Acid Gel Stain 10,000× (Molecular Probes).
9. Dark Reader Transilluminator (e.g., Clare Chemical Lab) or Safe Imager Blue Light Transilluminator (Molecular Probes).
10. Razor blade (sterile).
11. Gelase enzyme (1 U/µl) (Epicentre).
12. Thermomixer.
13. Ice bath.
14. 1.5 ml reaction tubes.
15. Centrifuge (refrigerated, 14,000 × *g*) for 1.5/2.0 ml tubes.
16. 3 M Sodium acetate (pH 7).
17. 100% Isopropanol p.A.
18. 70% EtOH.
19. TE buffer: 10 mM Tris–HCl (pH 7.5), 1 mM EDTA.
20. 0.8% agarose gel.
21. 0.5× TBE.
22. Fosmid Control (FC) DNA (included in EpiFOS Fosmid Library Production Kit).
23. Ethidium bromide.
24. UV-illuminator.
25. Aluminum foil.

2.1.4 Preparation of Epi100 Bacteria for Mass Transfection

1. EPI100 bacteria (included in EpiFOS Fosmid Library Production Kit).
2. 10 cm petri dishes.
3. LB broth: 10 g/l Bacto-Tryptone, 10 g/l NaCl, 5 g/l Yeast Extract.

4. LB agar: LB-Broth, 15 g/l Bacto-Agar.
5. 10 mM $MgSO_4$.
6. Autoclave.
7. Incubator (37 °C).
8. 50 ml Erlenmeyer flask.
9. UV-Spectrophotometer (Eppendorf).

2.1.5 Ligation of Size-Selected gDNA into pEpiFOS-Vector, Phage Packaging, and Testing Titer

1. Fast-Link Ligase and Buffer (included in EpiFOS Fosmid Library Production Kit).
2. 200 µl Microtubes.
3. 1.5 ml reaction tubes.
4. Thermomixer.
5. Sterile water.
6. Ice bath.
7. MaxPlax-Lambda Packaging Extract (included in EpiFOS Fosmid Library Production Kit).
8. Chloroform.
9. Epi100 bacteria (included in EpiFOS Fosmid Library Production Kit).
10. LB-Agar plates with 12.5 µg/ml chloramphenicol (10 cm petri dishes).
11. Incubator (37 °C).

2.1.6 Mass Transfection of E. Coli

1. Laminar flow hood.
2. LB broth with 10 mM $MgSO_4$ and 12.5 µg/ml Cloramphenicol.
3. 96-deep well plates.
4. Erlenmeyer flask.
5. EPI100 bacteria (included in EpiFOS Fosmid Library Production Kit).
6. Shaking incubator (37 °C).

2.1.7 Partitioning Fosmid Clones into Pools and Amplification of Fosmid Library

1. 96-deep well plates.
2. LB broth.
3. Multipette (Eppendorf).
4. Breathable seal.
5. Shaking incubator (37 °C).
6. LB Agar plates with 12.5 µg/ml Chloramphenicol.
7. Sterile glycerol (90%).
8. Eight-channel pipette.
9. Aluminum foil.
10. −80 °C freezer.

2.2 Isolation of Fosmid DNA from Fosmid Clone Pools

2.2.1 Plating and Scraping Fosmid Super-Pools

1. Large LB agar plates (22 cm × 22 cm) with 12.5 μg/ml Chloramphenicol.
2. LB broth, and LB broth with 12.5 μg/ml Cloramphenicol.
3. Inoculating loop.
4. Drigalski spatula.
5. 50 ml reaction tubes (Falcon).

2.2.2 Extraction of Fosmid Super-Pool DNA

1. Ice bath.
2. Centrifuge (12,500 × *g*) for 15 ml/50 ml Falcon tubes.
3. Centrifuge (refrigerated, 12,500 × *g*) for 1.5/2.0 ml tubes and 96-well plates.
4. QIAGEN Large-Construct Kit.
5. Fluted filters.
6. QIAGEN-tip 500 (to be purchased in addition, 2 per extraction required).
7. ATP disodium salt (AppliChem).
8. Water bath.
9. 100% Isopropanol p.A.
10. 70% EtOH.
11. 1.5 ml reaction tubes.
12. TE buffer: 10 mM Tris–HCl (pH 7.5), 1 mM EDTA.
13. Agarose gel electrophoresis system.

2.3 Fosmid Pool-Based Next Generation Sequencing Library Preparation

2.3.1 Barcoded Preparation of Fragment Libraries

1. Covaris S2 Sonicator (Covaris).
2. Covaris microTUBEs (Covaris).
3. Low TE buffer (Applied Biosystems).
4. Covaris G7 adaptor (Covaris).
5. Lonza Flash Gel System (Lonza).
6. 2.2% Lonza FlashGel Cassette (Lonza).
7. GeneRuler Low Range DNA Ladder, ready-to-use with 6× Orange DNA Loading Dye (Fermentas).
8. MinElute Gel Extraction Kit (QIAGEN).
9. 1.5 ml DNA LoBind tubes (Eppendorf).
10. End-It DNA End-Repair Kit (Epicentre).
11. Thermoshaker.
12. Qubit Fluorometric Quantitation (Thermo Scientific).
13. Qubit dsDNA HS Assay Kit (Thermo Scientific).
14. 4% Reliant NuSieve 3:1 agarose gel (Lonza).
15. UV Multibright Transilluminator (Intas).

16. SOLiD Fragment Library Oligo Kit (Applied Biosystems).
17. Quick Ligation Kit (NEB).
18. MinElute Reaction Cleanup Kit (QIAGEN).
19. Gel electrophoresis system (Bio-Rad).
20. Ethidium bromide (AppliChem).
21. DNA Polymerase I (*E. coli*) (10 U/μl) (NEB).
22. 100 mM dNTP-Mix (GeneAmp).
23. Microcentrifuge 5417R (Eppendorf).
24. Thermomixer.

2.3.2 Barcoded Preparation of Mate-Paired Libraries

1. SOLiD Mate-Paired Library Oligo Kit (Applied Biosystems).
2. HydroShear (Genomic Solutions, Inc.).
3. HydroShear Standard Shearing Assembly 1–5 kb (Genomic Solutions, Inc.).
4. QIAquick Gel Extraction Kit (QIAGEN).
5. End-It DNA End-Repair Kit (Epicentre).
6. 500 mM EDTA.
7. Quick Ligation Kit (NEB).
8. Microcentrifuge 5417R (Eppendorf).
9. 10× TAE (Applied Biosystems).
10. Agarose-LE (Applied Biosystems).
11. Gel electrophoresis system (any supplier).
12. Ethidium bromide (AppliChem).
13. 1 Kb DNA Ladder (Invitrogen).
14. Gel imaging system (any supplier).
15. Plasmid-Safe ATP-Dependent DNase (Epicentre).
16. DNA Polymerase I (*E. coli*) (10 U/μl) (NEB).
17. 100 mM dNTP Mix (GeneAmp).
18. T7 Exonuclease (10 U/μl) (NEB).
19. S1 Nuclease (400–1500 U/μl) (Invitrogen).
20. 3 M Sodium chloride, 5 M Sodium chloride.
21. Tris–HCl (500 mM, pH 7.5).
22. 1 M Magnesium chloride (Ambion).
23. Streptavidin Dynabeads, Dynal MyOne C1 (Thermo Fisher Scientific).
24. SOLiD Buffer Kit (including 1× Bead Wash Buffer, 1× Bind & Wash Buffer, 1× Low Salt Binding Buffer, Low TE Buffer, 1× TEX Buffer, 2-Butanol) (Applied Biosystems).
25. 100× BSA (NEB).

26. Six Tube Magnetic Stand (Applied Biosystems).
27. Vortexer (any supplier).
28. Rotator for 1.5–2.0 ml tubes (any supplier).

2.3.3 Large-Scale PCR of Fragment and Mate-Paired Libraries

1. PCR SuperMix (Invitrogen).
2. Cloned Pfu polymerase (2.5 U/μL) (Stratagene).
3. 4% Reliant NuSieve 3:1 agarose gel (Lonza).
4. GeneRuler Low Range DNA Ladder, ready-to-use with 6× Orange DNA Loading Dye (Fermentas).
5. MinElute Reaction Cleanup Kit (QIAGEN).
6. QIAquick Gel Extraction Kit (QIAGEN).
7. Gel electrophoresis system (any supplier).
8. Gel imaging system (any supplier).
9. SOLiD Mate-Paired Library Oligo Kit (Applied Biosystems).
10. SOLiD Fragment Library Oligo Kit (Applied Biosystems).
11. 96-Well GeneAmp PCR System 9700 (Applied Biosystems).
12. Lonza Flash Gel System (Lonza).
13. 2.2% Lonza FlashGel Cassette (Lonza).
14. Ethidium bromide (AppliChem).
15. 15 ml conical tubes (Falcon).
16. Microcentrifuge (12,500 × *g*).
17. Qubit Fluorometric Quantitation (Thermo Scientific).
18. Qubit dsDNA HS Assay Kit (Thermo Scientific).
19. 1.5 ml DNA LoBind tubes (Eppendorf).
20. Six Tube Magnetic Stand (Applied Biosystems).

2.3.4 Preparation of Fragment and Mate-Paired Sequencing Libraries for Emulsion PCR

1. Qubit Fluorometric Quantitation (Thermo Scientific).
2. Qubit dsDNA HS Assay Kit (Thermo Scientific).
3. SOLiD Buffer Kit (including 1x Bead Wash Buffer, 1× Bind & Wash Buffer, 1× Low Salt Binding Buffer, Low TE Buffer, 1× TEX Buffer, 2-Butanol) (Applied Biosystems).

2.4 Processing Next Generation Sequencing Libraries for Instrument Run

2.4.1 Emulsion PCR

1. 96-Well GeneAmp PCR System 9700 (Applied Biosystems).
2. SOLiD Buffer Kit (Applied Biosystems).
3. SOLiD ePCR Kit (Applied Biosystems).
4. 1 ml glass pipet (any supplier).
5. 5 ml glass pipet (any supplier).
6. ULTRA-TURRAX Tube Drive (IKA).
7. SOLiD ePCR Tubes and Caps (IKA).

8. Covaris S2 Sonicator (Covaris).
9. Covaris-2 Series Machine Holder for 1.5-ml microcentrifuge tube (Covaris).
10. Covaris-2 Series Machine Holder for 0.65-ml microcentrifuge tube (Covaris).
11. 15 ml tube (Falcon).
12. 50 ml tube (Falcon).
13. Vortexer (any supplier).
14. Semi-automated Xstream pipettor (Eppendorf).
15. Repeater plus pipette (Eppendorf).
16. MicroAmp Optical 96-Well Reaction Plates (Applied Biosystems).
17. MicroAmp Optical Adhesive Film (Applied Biosystems).
18. Six Tube Magnetic Stand (Applied Biosystems).
19. NanoDrop ND1000 Spectrophotometer (Thermo Scientific).
20. Nuclease-free water.

2.4.2 Breaking the Emulsion PCR

1. SOLiD Buffer Kit (Applied Biosystems).
2. Repeater plus pipette (Eppendorf).
3. 50 ml tube (Falcon).
4. SOLiD Emulsion Collection Tray Kit (Applied Biosystems).
5. Fume hood (any supplier).
6. NanoDrop ND1000 Spectrophotometer (Thermo Scientific).
7. Six Tube Magnetic Stand (Applied Biosystems).
8. 1.5 ml LoBind Tubes (Eppendorf).

2.4.3 Enrichment of Templated Beads

1. SOLiD Bead Enrichment Kit (Applied Biosystems).
2. SOLiD Buffer Kit (Applied Biosystems).
3. Microcentrifuge (12,500 × *g*).
4. 15 ml tubes (Falcon).
5. Six Tube Magnetic Stand (Applied Biosystems).
6. Covaris S2 Sonicator (Covaris).
7. Covaris-2 Series Machine Holder for 1.5-ml microcentrifuge tube (Covaris).
8. Covaris-2 Series Machine Holder for 0.65-ml microcentrifuge tube (Covaris).
9. 0.5 ml LoBind Tubes (Eppendorf).
10. 1.5 ml LoBind Tubes (Eppendorf).
11. 2.0 ml LoBind Tubes (Eppendorf).

2.4.4 3′-End Modification of Enriched Templated Beads

1. Six Tube Magnetic Stand (Applied Biosystems).
2. SOLiD Buffer Kit (Applied Biosystems).
3. Covaris S2 Sonicator (Covaris).
4. Covaris-2 Series Machine Holder for 1.5-ml microcentrifuge tube (Covaris).
5. SOLiD Bead Deposition Kit (Applied Biosystems).
6. 1.5 ml LoBind Tubes (Eppendorf).
7. NanoDrop ND1000 Spectrophotometer (Thermo Scientific).

2.4.5 Bead Deposition on SOLiD Sequencing Slide and Instrument Run

1. SOLiD Bead Deposition Kit (Applied Biosystems).
2. Covaris S2 Sonicator (Covaris).
3. 1.5 ml LoBind Tubes (Eppendorf).
4. Six Tube Magnetic Stand (Applied Biosystems).
5. Covaris-2 Series Machine Holder for 1.5-ml microcentrifuge tube (Covaris).
6. SOLiD Slide Kit (Applied Biosystems).
7. SOLiD Bead Deposition Kit (Applied Biosystems).
8. SOLiD Fragment Library Sequencing Kit (Applied Biosystems).
9. SOLiD Mate-Paired Library Sequencing Kit (Applied Biosystems).
10. SOLiD Instrument Buffer Kit (Applied Biosystems).
11. SOLiD Deposition Chamber 1, 4, 8 well (Applied Biosystems).

2.5 Computational Analysis of Fosmid Sequences and Haplotype Assembly

The computational requirements for primary and secondary sequence analysis include high-end computing servers with > 1 Terabyte (TB) disk space (RAID system) and > 32 Gigabyte (GB) Memory (RAM), and > 4 CPU cores per server or cluster node.

2.5.1 Hardware Requirements

2.5.2 Software Requirements

The following list of software tools is used at different (generally successive) stages of this process:

For SOLiD NGS Analysis: BioScope™ Software

For other NGS systems such as Illumina:

Bwa [13]: http://bio-bwa.sourceforge.net/

Bowtie2 [14]: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

Picard: http://broadinstitute.github.io/picard/

NGSEP [15]: http://sourceforge.net/projects/ngsep/

GATK [16]: https://www.broadinstitute.org/gatk/

Samtools [17]: http://samtools.sourceforge.net/

Fosmid detector [3] and ReFHap [5, 11]: http://www.molgen.mpg.de/~genetic-variation/SIH/Data/algorithms. This address includes a README.txt file with detailed instructions to run both the fosmid detector and ReFHap.

3 Methods

The following procedures apply to haplotype-resolving one individual genome.

3.1 Fosmid Library Construction

To establish a high quality individual fosmid library, two kits proved to work well and efficiently in conjunction, the DNA Extraction Kit from Stratagene and the EpiFOS Fosmid Library Production Kit from Epicentre. The first one, specifically, provides three solutions almost ready to use, and a protease mixture to digest cellular proteins and RNase to eliminate RNA. Its protocol is applicable to isolate HMW gDNA from blood, whole tissue, and cultured cells. The second kit can be applied to establish a fosmid library from sheared gDNA, and provides all components and reagents required to end-repair and ligate the sheared gDNA into a single copy fosmid vector, phage-package the fosmid clone DNA, and transfect *E. coli* cells (*see* **Note 1**).

3.1.1 High Molecular Weight Genomic DNA Isolation

The preparation of fosmid clones requires very high quality HMW gDNA.

1. Start with 8 ml EDTA blood in a 50 ml reaction tube for the extraction of a total of 40–50 μg HMW gDNA per individual sample. Add 42 ml 1× Solution 1 according to the manufacturer's protocol, and incubate the sample on ice for 2 min, followed by spinning it at 350 × *g* and 4 °C for 15 min. Discard the supernatant carefully, because the pellet is very instable. Resuspend the pellet in 11 ml of Solution 2, add 5 μl pronase (225 mg/ml) to a final concentration of 100 μg/ml and incubate the sample in a shaking water bath for 1 h at 60 °C. Transfer it on ice for 10 min, add 4 ml of Solution 3, and invert the tube several times (*see* **Note 2**). Incubate the sample for 5 more min on ice to precipitate the cellular proteins. Spin the tube at 2000 × *g* and 4 °C for 15 min and transfer the supernatant to a sterile 50 ml tube by using a large-bore pipet tip (*see* **Note 3**).
2. For RNA digestion, add RNase (10 mg/ml) to a final concentration of 20 μg/ml and incubate the sample in a water bath for 15 min at 37 °C.
3. The HMW gDNA can be precipitated after adding two volumes of absolute EtOH and gently inverting the tube. Use a sterile glass rod to spool the gDNA and rinse it with 70% EtOH. Dry the spooled DNA, transfer it into a sterile 50 ml tube, and carefully resuspend it in 500 μl pre-warmed TE buffer (*see* **Note 4**). Avoid any vortexing or pipetting, which might degrade the HMW gDNA.

4. Dissolve the pellet on a slowly rotating tube roller at 4 °C (*see* **Note 5**). Calculate the yield and concentration of your sample by measuring the OD_{260}. Store the DNA at 4 °C.
5. Control the quality of the extracted HMW gDNA: run a Pulsed Field Gel Electrophoresis (PFGE) with a 1% agarose gel in 0.5× TBE, load 3 µl of the HMW gDNA, use λ DNA-Mono Cut Mix DNA Ladder in the outer lanes as reference, run PFGE at 6 V/cm, switch time 0.2–2 s, 13–16 h (*see* **Note 6**).

3.1.2 Mechanical Shearing of HMW gDNA and End-Repair

To provide the basis for the fosmid library construction, the extracted HMW gDNA needs to be fragmented into ~40 kb segments and end-repaired to be cloned into a fosmid vector. DNA fragments larger than ~60 kb or smaller than ~20 kb can prevent phage-packaging at a later stage. Moreover, the use of fragments <20 kb might result in the formation of chimeric clones.

1. Dilute 8 µg HMW gDNA to a final concentration of 20 ng/µl.
2. Shear the DNA using a HydroShear shearing device; use the 4–40 kb (LARGE) shearing assembly. Since every shearing assembly has slightly different shearing properties, test different speed codes at the outset: prepare three aliquots of 8 µg DNA (400 µl) and test them with speed codes "16," "17," or "18," retraction speed "20," and 25 shearing cycles. Check and compare the results on an agarose gel.
3. Alternatively, if a HydroShear device is not available, the DNA can be sheared manually using a sterile 1 ml disposable syringe with a 23G 1¼ needle, aspirating 400 µl of the diluted HMW gDNA and pulling the syringe up and down for 50 s (12 times) (*see* **Note 7**).
4. After controlling the shearing results on by PFGE (*see* Subheading 3.1.1, **step 5**), precipitate the sheared DNA with isopropanol (100%) (*see* Subheading 3.1.3, **step 5**) and resuspend the sheared gDNA in 26 µl TE.
5. To produce blunt-end gDNA fragments, mix 4 µl End-Repair 10× Buffer, 4 µl of 2.5 mM dNTPs and 4 µl of 10 mM ATP into a 1.5 ml tube, add 26 µl of sheared gDNA and 2 µl End-Repair Enzyme (final volume 40 µl), incubate the mix for 60 min at room temperature, and transfer the tube into a preheated thermomixer (70 °C) for 10 min to inactivate the enzymes (*see* **Note 8**).

3.1.3 Size Selection of End-Repaired gDNA Fragments by PFGE

Size-selection by PFGE is performed to guarantee suitable DNA fragments for library production.

1. Prepare a 1% low melting point (LMP) agarose gel (with 200 ml 0.5× TBE), using a preparative comb with four slots (*see* **Note 9**). Mix 40 µl end-repaired gDNA with 7 µl 6× loading dye, and slowly pipet gDNA sample into one preparative

slot. Keeping one well empty between gDNA sample and size ladder, load 1 μl λ DNA-Mono Cut Mix DNA ladder (mixed with 1 μl 6× loading dye and 4 μl dH_2O) on both sides of the gDNA sample. Run a PFGE at 6 V/cm; initial sweep time 0.5; final sweep time 2.0 s; for 20–22 h.

2. Use Sybr Gold Dye to stain the agarose gel to avoid the need for UV exposure. Prepare a staining solution by pipetting the 10 μl Sybr Gold Stock Solution into 100 ml of 0.5× TBE (in a 100 ml glass flask wrapped in aluminum foil), pour the staining solution into a plastic tray, lift the gel into the staining bath, and leave for at least 30 min (*see* **Note 10**).
3. Visualize stained gDNA fragments on a Dark Reader Transilluminator. Ensure that the bulk of sheared gDNA has migrated within the correct size range (20–50 kb). Make a hole with a pipet tip at the 30 kb and 48 kb band of the size ladder, turn off illumination, use a sterile razor blade to select sheared gDNA by making a cut parallel to the 30 and 48 kb size ladder band (using the holes as marks) and excising the gel pieces with size-selected DNA fragments by vertical cuts (*see* Fig. 2 and **Note 11**).
4. To recover the size-selected gDNA, warm Gelase 50× Reaction Buffer in a thermomixer at 45 °C, set a second thermomixer (or a water bath) to 70 °C, push excised gel slice into a 2 ml tared tube, and weigh the gel piece. Translate solid gel weight into volume of molten agarose (1 mg of solid gel will result in 1 μl molten agarose), calculate volume of 50× Gelase Buffer required to yield a 1× Buffer, and calculate Gelase enzyme units (1 U/μl; 1 U Gelase enzyme per 100 μl molten agarose) needed after the subsequent step. Heat tube with weighed gel

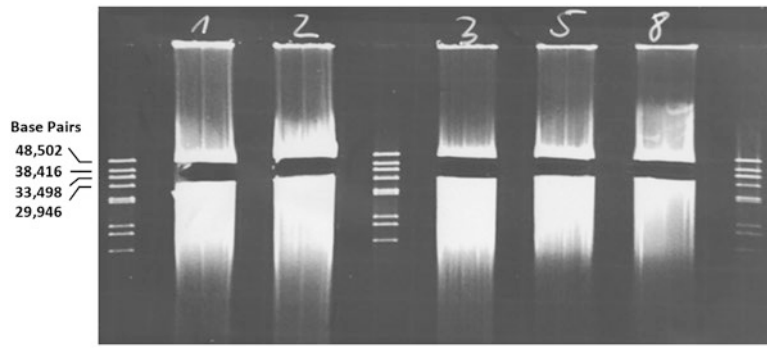

Fig. 2 Size-selection of end-repaired gDNA fragments after PFGE. Sheared gDNA from five different individuals resolved by PFGE on a 1 % low melting point (LMP) agarose gel (200 ml 0.5× TBE) together with a λ DNA-Mono Cut Mix DNA Ladder; gel stained with Sybr Gold Dye. Gel pieces with fragments between ~30 and 48 kb in size were cut out

slices in a 70 °C thermomixer (or water bath), keep for 10–15 min until the agarose is molten, quickly transfer tubes with molten agarose to 45 °C thermomixer, add appropriate volume of pre-warmed 50× Gelase Buffer and appropriate units of Gelase enzyme, gently mix the solution, and incubate for at least 60 min at 45 °C. Heat inactivate Gelase enzyme at 70 °C, then chill the mix for 15 min in an ice bath. Centrifuge the reaction mix for 20 min at maximum speed (11,000 × *g*), and transfer most of supernatant to a new 1.5 ml tube, making sure to not aspirate the gelatinous pellet.

5. Purify size-selected gDNA after gelase digest, add 1/10 volume 3 M sodium acetate (pH 7), mix gently, add 1 volume isopropanol, gently invert the tube, leave for 10 min at room temperature, centrifuge for 20–30 min at full speed, remove supernatant and wash pellet with 1 ml 70% ethanol twice, air-dry pellet for 10 min, and resuspend in 11 μl TE.
6. To control for quality and quantity of size-selected gDNA, run an 0.8% TBE agarose gel (minigel of 7–10 cm length), load 1 μl of size-selected gDNA, use 1 μl Fosmid Control (FC) DNA from the EpiFOS Fosmid Library Production Kit (Epicentre) (size 40 kb, 100 ng/μl) on both sides of the sample, additionally load 1 μl each of a 1:2 and 1:4 dilution of the FC DNA, run the gel at 6 V/cm for 40–60 min, stain with ethidium bromide, and visualize on an UV-illuminator and check the quantity and size of the size-selected gDNA by comparing the sample to the FC DNA dilutions.

3.1.4 Preparation of Epi100 Bacteria for Mass Transfection

The EpiFOS Fosmid Library Production Kit (Epicentre) provides a glycerol stock of EPI100 bacteria suitable for mass transfection.

1. Use multiple 10 cm petri dishes, prepare LB agar [without chloramphenicol (CA)], autoclave and pour warm into dishes, and use solidified LB agar plates to streak Epi100 plating strain (*see* **Note 12**). Incubate at 37 °C overnight, seal and store the plate at 4 °C. The day before the phage-packaging procedure, pick a single colony from the plated strain and inoculate a 50 ml Erlenmeyer flask with fresh LB broth and 10 mM $MgSO_4$. Shake the flask overnight at 37 °C, and use 2 ml of the overnight culture for the next step (Subheading 3.1.4, **step 2**). The overnight culture can be stored at 4 °C (*see* **Note 13**).
2. On the day of phage packaging, prepare an Erlenmeyer flask with fresh 50 ml LB broth and 10 mM $MgSO_4$, then inoculate with 2 ml of the Epi100 overnight culture from **step 1**; shake at 37 °C until OD_{600} reaches 0.85, measuring every 20 min.
3. Prepare multiple LB plates with CA in 10 cm petri dishes; these will be used to determine the titer of packaged fosmid clones (phage particles) in Subheading 3.1.5. Seal and store LB plates at 4 °C (*see* **Note 14**).

3.1.5 Ligation of Size-Selected gDNA into EpiFOS-Vector, Phage Packaging, and Testing Titer

In order to establish an individual fosmid library that ensures a ~7× coverage of the haploid and ~15× coverage of the diploid genome, a total of 1.44×10^6 fosmid clones need to be generated.

1. For the ligation reaction, combine 1 μl Fast-Link Ligation 10× Buffer, 1 μl 10 mM ATP, 1 μl pEpiFOS-Vector (500 ng), and 250 ng of the size-selected gDNA (corresponding to a 10:1 M ratio of vector:insert) in a new 200 μl microtube, add 1 μl Fast-Link Ligase (2 U/μl) and dH_2O to a final reaction volume of 10 μl, seal the tube and incubate the reaction for 2 h at room temperature, inactivate by placing micro tube at 70 °C for 10 min. In order to control the fosmid library production procedure, prepare a control reaction, use 2.5 μl of 40 kb FC insert DNA provided in the kit instead of size-selected gDNA. The control ligation reaction used to determine the packaging efficiency (*see* Subheading 3.1.5, **step 3**) should yield more than 1×10^7 colony forming units (cfu)/ml of control insert DNA.
2. For phage packaging, thaw one tube of MaxPlax-Lambda Packaging Extract on ice, immediately transfer half of the extract (25 μl) into a new 1.5 ml tube, and store the remaining half of packaging extract at −80 °C freezer. Add 10 μl of ligated DNA and mix by pipetting the reaction repeatedly (do not introduce air bubbles), spin down the tube, and incubate the reaction mixture for 90 min in a thermomixer at 30 °C. Five min before the end of the incubation period, take the remaining half of the packaging extract out of the freezer and put it on ice to thaw it. Then add it to the reaction mixture and incubate the reaction again for 90 min at 30 °C. Finally, add 940 μl Phage Dilution Buffer (PDB) to a total volume of 1 ml, mix gently, then add 25 μl of chloroform to the reaction, and store at 4 °C.
3. Phage titer testing: To determine the number of phage particles that contain fosmid clones, use 10 μl of the 1 ml phage packaging reaction (*see* **step 2**), add 990 μl PDB to generate a $1{:}10^2$ phage dilution, then use 100 μl of the $1{:}10^2$ dilution and 900 μl of PDB to prepare a $1{:}10^3$ dilution. Processing two dilutions in parallel is useful to be able to determine the efficiency of the phage packaging reaction as precisely as possible. Then mix 10 μl of each phage dilution with 100 μl of prepared Epi100 bacteria (*see* Subheading 3.1.4) into a 1.5 ml tube each, incubate both mixtures for 20 min at 37 °C by the use of a thermomixer, then plate the total volumes of 110 μl of transfected Epi100 cells each on an LB plate with CA (petri dish), and incubate at 37 °C overnight. To make sure that the packaging efficiency is sufficient to establish a complex fosmid library, count the numbers of colonies grown from each of the phage dilutions to calculate the total expected number of circular fosmid clones packaged by phages (*see* **step 2**) as follows: (Number of colonies x dilution factor × 1000 μl) divided by

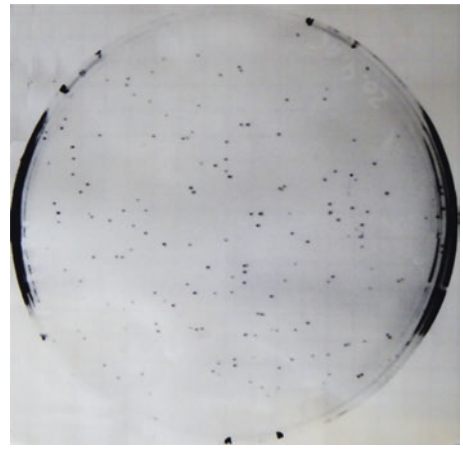

Fig. 3 Titer testing to determine the total number of fosmid clones expected from mass transfection. A petri dish (agar plate with chloramphenicol) with multiple bacterial colonies is shown; these were grown overnight from Epi100 bacteria transfected with a $1{:}10^2$ phage dilution. 166 colonies are counted, allowing calculation of the total number of expected fosmid clones to nearly 1.7×10^6, sufficient to generate an individual fosmid library from a single cloning procedure

(volume of plated phage particles in µl) (*see* **Note 15** and Fig. 3). If the expected total number of fosmid clones is below 1.44×10^6—that is, if less than 150 fosmid clones are counted from plating the $1{:}10^2$ phage dilution—start a second fosmid cloning round beginning with the first step, shearing of 8 µg HMW gDNA. Store the packaged phages in PDB at 4 °C until the required number of fosmid clones has been obtained.

3.1.6 Mass Transfection of E. Coli

Make sure that all the following steps are performed in one workflow under sterile conditions. Specifically, work under a sterile laminar flow hood and use sterile media only.

1. To begin with, prepare 440 ml LB broth with 10 mM $MgSO_4$ and 12.5 µg/ml CA, aliquot the prepared LB into 3×96-deep well plates, 1.5 ml per well.
2. For mass transfection, prepare all phage packaging reactions required to generate 1.5×10^6 fosmid clones and determine their total volume (*see* Subheading 3.1.5, **step 3**). Fill an Erlenmeyer flask with the prepared Epi100 bacteria (*see* Subheading 3.1.4) to bring the final transfection volume up to 30 ml. For example, if one phage packaging reaction of 1 ml has proven sufficient to achieve a total of 1.5×10^6 phages, start with 29 ml of Epi100 cells.
3. Add the phages to the Epi100 cells and let the phage particles adsorb for 20 min at 37 °C. Preserve an aliquot of 20 µl from the mass transfection mix, expected to contain ~1000 fosmid clones, in order to control the clone pool complexity; store on ice until use in Subheading 3.1.7, **step 2**.

3.1.7 Partitioning Fosmid Clones into Pools and Amplification of Fosmid Library

Partitioning the fosmid clones into 3×96-deep well plates, i.e., 288 wells, allows generation of "haploid clone pools" that contain 5000 fosmid clones per well. This library format minimizes the probability that complementary parental haplotypes co-occur in one well. Thus, multiple fosmid pools, each representing a random mixture of ~40 kb haploid DNA segments covering 5% of the haploid genome, can be sequenced to saturate both haploid genomes. The partitioning and amplification of an entire fosmid library with agar plates is very labor-, and cost-intensive. To simplify the procedure we have chosen to use liquid cultures at this stage as described below (*see* **Note 16**).

1. Dispense 100 µl of the mass transfection into each well of the 3×96-deep well plates, expected to correspond to 5000 fosmid clones per well, and cover with breathable seal. Each well contains prepared LB broth (*see* Subheading 3.1.6, **step 1**).
2. After partitioning fosmid clones into deep wells, incubate for 20 h at 37 °C at 200 rpm in a shaking incubator to amplify the clones.
3. Take the 20 µl sample preserved in Subheading 3.1.6, **step 3**, and plate 5 µl (expected to correspond to ~250 colonies) and 2.5 µl (expected to correspond to ~125 colonies) each on a 10 cm LB plate with CA and incubate at 37 °C overnight.
4. The next day, count the number of grown clones. Determine the obtained clone pool complexity as follows: (Number of cfu/Volume of plated fosmid clones) = (Number of cfu per well/Dispensed mass transfection volume). For example, if 260 cfu are counted on an LB plate, the number of cfu, or fosmid clones that have been obtained per well in the library is calculated as (260 cfu/5 µl) = (Number of cfu per well/100 µl), and is 5200 in this case
5. For long-term storage of fosmid libraries, aliquot 250 µl of glycerol (90%) into each of the wells of the 3×96-well plates to a final volume of 1.85 ml, mix and store at −80 °C..
6. In order to prepare working plates, thaw glycerol stock deep well plates and pipet 150 µl per well from the deep well stock plate into a 96-well working plate; make sure to pipet deep well content up and down at least once before removing 150 µl to capture all individual clones. Use a plate with wells of at least 250 µl of capacity.
7. To prepare super-pool plates with 15,000 fosmid clones per well, combine the three 96-well working plates into a single 96-well plate. Transfer 50 µl from each 96-well working plate into a single, new 96-well plate; use an 8-channel pipette to always combine the wells with identical positions on the working plates. Cover the super-pool plate with breathable seal, pack into aluminum foil and store in −80 °C freezer.

3.2 Isolation of Fosmid DNA from Fosmid Clone Pools

Due to the usage of the EpiFos single-copy vector, each transfected Epi100 bacterium will only contain one fosmid clone. To isolate sufficient amounts of fosmid DNA from single fosmid super-pools for sequencing library preparation, while preserving library complexity, the use of LB agar plates for this amplification step (*see* **Note 16**) has proven to be indispensable. As a consequence of such large-scale cultivation of Epi 100 cells, large amounts of Epi100 genomic DNA and proteins must be removed, requiring additional purification steps. The fosmid clones scraped from the agar plate are divided into two portions for more efficient exonuclease digest of bacterial genomic DNA, and the proteins are filtrated before extraction of fosmid insert DNA (*see* **Note 17**).

3.2.1 Plating and Scraping Fosmid Super-Pools

As outlined in the Introduction, routinely, 40–48 super-pools per fosmid library are sequenced to be able to phase approximately 85–92 % of the heterozygous SNPs. Thus, this number of super-pools was amplified for subsequent DNA isolation, preparing two plates per super-pool.

1. Prepare two large (22×22 cm) LB agar plates per super-pool with CA. Thaw 96-well super-pool plates with 15,000 fosmid clones per well on ice. Scratch thawed fosmid-glycerol mixture with the inoculating loop to yield approximately 3 μl fosmid-glycerol mixture per well, and mix with 1 ml LB medium and CA in a 1.5 ml tube. Pipet 500 μl of the LB-clone mix onto one LB plate and the other half of the LB-clone mix onto the second LB plate, spread with Drigalski spatula, and incubate at 37 °C overnight.
2. To scrape the clones, rinse 3×5 ml LB medium over the first incubated LB plate, carefully scrape colonies from agar plate with a Drigalski spatula, and pipet clones from LB plate into a single 50 ml Falcon tube. Repeat for the second LB plate containing the second half of the super-pool, and pipet collected clones into a new 50 ml Falcon tube.

3.2.2 Extraction of Fosmid Super-Pool DNA

1. Chill centrifuge to 4 °C, and centrifuge the two 50 ml Falcon tubes each containing half of the clones from a single super-pool at 6000 ×*g* for 15 min. For extraction, the QIAGEN Large-Construct Kit is used, which contains most of components, reagents, and buffers. Resuspend pellet in 10 ml P1 Buffer. Add 10 ml P2 Buffer and mix by inverting rigorously (5×) to lyse cells (but do not vortex), leave reaction for 2 min at room temperature, stop cell lysis reaction by adding 10 ml prechilled (4 °C) P3 buffer, and cool reaction 10 min on ice. Centrifuge the Falcon tubes for 30 min at 4 °C at a minimum of 12,500×*g* (*see* **Note 18**).
2. Prepare two fluted filters, wet the filter paper with dH_2O and put each filter onto a new 50 ml Falcon tube, pipet the supernatant (containing fosmid DNA) from the first centrifuged

Falcon tube onto one pre-wetted fluted filter, and the supernatant from the second Falcon tube onto a second pre-wetted fluted filter, wait until supernatant has been filtrated.

3. Equilibrate two QIAGEN-tips 500 (from the QIAGEN Large-Construct Kit) with 10 ml QBT Buffer, transfer the clear filtrate from each Falcon tube to one QIAGEN-tip, wash the tip twice with 30 ml QC Buffer, elute fosmid super-pool DNA with 15 ml preheated (65 °C) QF Buffer. Pool the eluates from both Falcon tubes, which are from a single fosmid super-pool.
4. Precipitate fosmid super-pool DNA by adding 0.7 volumes (21 ml) isopropanol, mix and centrifuge for 40 min at 12,500 × *g*. Wash pellet with 5 ml ethanol (70%), centrifuge for 30 min at 12,500 × *g*, air-dry pellet at room temperature, and redissolve DNA in 9.5 ml EX Buffer.
5. To remove contaminating genomic DNA from the EPI100 bacteria and nicked or damaged large-construct DNA, an exonuclease digestion is strongly recommended. Therefore, prepare first 100 mM ATP solution: Dissolve 2.75 g ATP (adenosine 5′triphosphate disodium salt, anhydrous, MW 551.14) in 40 ml distilled water, adjust the pH to 7.5 with 10 M NaOH, bring the volume up to 50 ml, distribute into 300 µl aliquots, and store at −20 °C. Secondly, prepare ATP-dependent Exonuclease by resuspending 80 µg ATP-Dependent Exonuclease in 225 µl Exonuclease Solvent. Then add 200 µl ATP-dependent Exonuclease and 300 µl of the 100 mM ATP solution to dissolved DNA to remove noncircular DNA from the bacterial genome, mix thoroughly, and incubate reaction mix at 37 °C overnight.
6. Equilibrate one QIAGEN-tip 500 with 10 ml QBT buffer, mix the exonuclease-digested fosmid super-pool DNA with 10 ml QS buffer, and transfer the whole DNA mix to QIAGEN-tip 500. After the DNA has passed the tip, wash two times with 30 ml QC Buffer, elute and precipitate fosmid super-pool DNA as described in **steps 3** and **4**, resuspend air-dried DNA pellet in 300 µl TE, and dissolve at room temperature overnight. Transfer fosmid DNA into a new 1.5 ml Eppendorf tube and rinse Falcon tube with additional 100 µl TE to collect all fosmid super-pool DNA. Check the quality and quantity of extracted fosmid super-pool DNA on a 1.5% agarose gel, run at 120 V for 20 min (*see* Fig. 4).

3.3 Fosmid Pool-Based Next Generation Sequencing Library Preparation

At the time of developing and applying our method to production, we used a SOLiD System from ABI/Life Technologies. The extracted super-pool fosmid DNA can also be analyzed in conjunction with any other NGS technology following the manufacturer's protocols for NGS sequencing library preparation. All reagents have been provided with the SOLiD system sequencing kits, if not

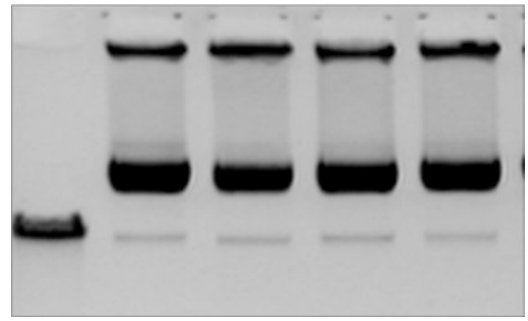

Fig. 4 Fosmid super-pool DNA controlled after extraction. The DNA isolated from four super-pools is shown resolved on a 1.5 % agarose gel stained with ethidium bromide; on the *left*, (linearized) fosmid control (FC) DNA, the faint DNA bands of the same size as FC DNA seemingly represent linear fosmid DNA, which has escaped purification. The larger, thick bands represent the circular fosmid DNA

specified otherwise. The fosmid DNA isolated from a single fosmid super-pool is used as input to prepare a sequencing library; thus, up to 96 sequencing libraries could be generated if one would want to utilize all super-pools from an individual fosmid pool library. To allow parallel processing of multiple NGS sequencing libraries, we created barcode tags by modification of the P1-adaptor sequence. With the barcoded sequences, up to 16 super-pools could be multiplexed later in the clonal amplification step and the sequencing run.

3.3.1 Barcoded Preparation of Fragment Libraries

In this case, the isolated fosmid super-pool DNA is used as input to create a library of short DNA sequencing template molecules 100–150 bp in length.

1. At least 30 min ahead of time, prepare Covaris S2 Sonicator for DNA shearing. Fill with deionized water to mark level 12 (water should cover glass microtube), degas for 30 min while continuously keeping the pump on, then cool down to 4 °C. Make sure that no bubbles appear at the bottom of the tube after degasing has finished (*see* **Note 19**).
2. Use 3 μg of extracted fosmid pool DNA in 100 μl Low TE buffer, place capped Covaris microtube into Covaris G7 adaptor. Transfer 100 μl DNA into microtube using extra thin pipet tip to avoid damaging of microtube presplit septa (cap of Covaris microtube) and make sure to not introduce an air bubble at the bottom of the microtube; shear the input DNA for 360 s ("DC" 20 %, intensity 5, cycles burst 200, time 60 s, six cycles, frequency sweeping mode) to generate fragments between 100 and 150 bp in size.
3. Control the shearing result by running a Lonza FlashGel System with a 2.2 % Lonza FlashGel Cassette: mix 3 μl of sonicated DNA with 3 μl 1× Loading Dye (diluted from 6× Orange DNA Loading Dye), pipet 2.5 μl Low Range DNA Ladder in

the first and last lane, run gel at 275 V for 6 min, visualize DNA while migrating. If the bulk of sheared DNA is above a size range of 100–150 bp, repeat Covaris shearing as described above for 180 s, and control fragment sizes again on a Lonza gel. Concentrate sheared DNA with QIAGEN MinElute columns (from MinElute Gel Extraction Kit, QIAGEN) as per manufacturer's protocol, elute twice in 21 μl elution buffer, take 1.5 μl of eluted sample to measure final DNA concentration, and transfer sheared DNA to a 1.5 ml LoBind tube.

4. End-repair of fragments for blunt-end ligation: mix 40 μl of sheared fosmid pool DNA, 6 μl End-Repair Buffer, 6 μl 10× dNTPs, 6 μl ATP, 2 μl End-Repair enzyme from the End-It DNA End-Repair Kit (Epicentre), incubate at 21 °C in a thermoshaker for 30 min, purify with MinElute Columns as per protocol, elute twice in 20 μl elution buffer, quantify with Qubit per manufacturer's protocol.
5. Size selection of end-repaired fragments: run a 4% Reliant NuSieve 3:1 agarose gel with ethidium bromide (*see* **Note 20**), mix total volume of end-repaired DNA with 7 μl Loading Dye (6×), load three slots each with 15 μl DNA, keep one well on both sides of the samples empty, use 5 μl Low Range DNA ladder at both sides of the gel, run for 40 min at 110 V (210 mA). Visualize sample on an UV Multibright Transilluminator shortly, use UV protection shield to excise the DNA between bands 100 and 150 bp (*see* Fig. 5a), transfer gel piece into one 2 ml tube and cut into smaller pieces, weigh

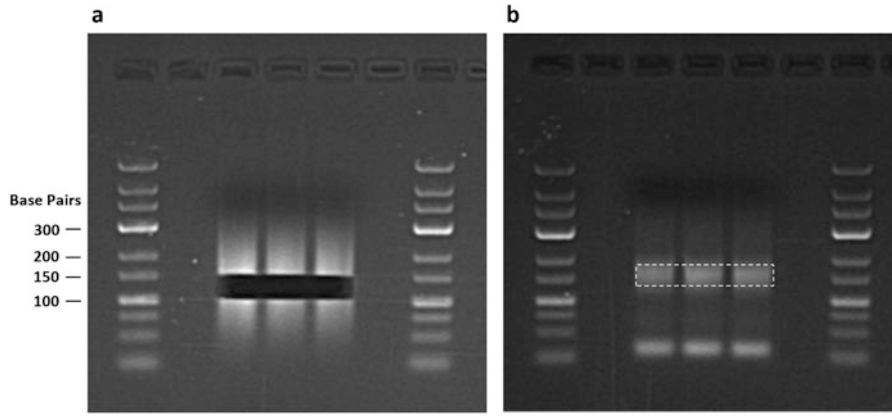

Fig. 5 Size-selection of sheared fosmid super-pool DNA for preparation of fragment sequencing libraries. (**a**) Three aliquots of sheared, end-repaired fosmid clone DNA fragments before ligation of adaptor sequences are shown resolved on a 4% Reliant NuSieve 3:1 agarose gel (Lonza) with ethidium bromide together with a Low Range DNA Ladder, and a gel piece containing DNA fragments of 100–150 bp in size has been cut out. (**b**) The excised DNA fragments are shown resolved after ligation of adaptor sequences, and are now slightly larger, in the range of 150–200 bp in size. These ligation products are again excised, as indicated by the *white dashed lines*, for subsequent purification and nick translation. The bands at the bottom represent excess adaptor sequences that were not ligated

tared tube and digest gel as described by the manufacturer's protocol, *see* also Subheading 3.1.3, **step 4**; then purify with QIAGEN MinElute columns, elute twice in 25 μl elution buffer, and use 1 μl to determine DNA concentration with Qubit (*see* **Note 21**).

6. The barcoding of a fosmid super-pool DNA sample is achieved by ligating a modified P1 adaptor sequence containing a unique barcode sequence tag of 6 bp (*see* **Note 22** including a list of these unique barcode sequences) to the DNA fragments instead of the universal adaptors provided by the SOLiD Fragment Library Oligo Kit. First, the necessary picomoles (pmol) of adaptor required for ligation are determined according to the following formula:

   1 μg DNA × 10^6 pg/μg × 1 pmol/660 pg × 1/(Average insert fragment size) = 17 pmol

   (# μg DNA) × (17 pmol DNA) = (# pmol DNA for adaptor ligation)

   (# pmol DNA for adaptor ligation) × (30) = (# pmol adaptors needed)

   (# pmol adaptors needed)/(# pmol/μl stock) = (# μl adaptor needed)

   With an average size of 125 bp for sheared DNA fragments, the volume of universal adaptors can be calculated as follows: (Total amount of purified DNA in pg/660)/125 × (30/50).

7. Ligation reaction: pipet 47–49 μl of purified fosmid pool DNA into a 1.5 ml LoBind tube, add 100 μl 2× Quick Ligation Buffer (from the Quick Ligation Kit, NEB), P1 and P2 adaptors (50 pmol/μl) as calculated in **step 6**, and 5 μl Quick Ligase, fill with dH_2O to a total reaction volume of 200 μl, incubate exactly 10 min in a thermomixer preheated to 21 °C, and stop the reaction instantly by adding 600 μl ERC buffer. Purify the reaction with MinElute columns, using the QIAGEN MinElute Reaction Cleanup Kit according to the manufacturer's protocol, and elute twice in 20 μl elution buffer to a final sample volume of 40 μl.

8. Size-selection of ligation products: mix the total sample volume with 7 μl 6× Loading Dye, load sample in three aliquots of 16 μl into three wells of a Reliant 4% NuSieve 3:1 agarose gel (with ethidium bromide), keep one well on each side empty, use Low Range DNA Ladder on both sides of gel, run for 40 min at 110 V, visualize on UV Transilluminator, size-select as described earlier (Subheading 3.3.1, **step 5**) the band between 150 bp and 200 bp (size corresponds to fosmid pool DNA plus ligated adaptor sequences; *see* Fig. 5b), and weigh gel slices in tared 2.0 ml tube. If gel slice weighs more than 400 mg, divide gel pieces into two tubes (*see* **Note 23**).

9. Purification of size-selected ligation products with MinElute Gel Extraction Kit (QIAGEN) according to the manufacturer's protocol: add 6 volumes QG buffer to 1 volume (equivalent to 100 mg or 100 µl) of gel, dissolve the gel by vortexing the tube at room temperature, vortex tube every 2–3 min for minimally 15 min until gel has completely dissolved. Add 1 gel volume of isopropanol, invert the sample, apply to MinElute column and centrifuge for 1 min, discard flow-through, add 500 µl QG buffer and centrifuge 1 min, discard flow-through, wash with 750 µl PE buffer, centrifuge 1 min, discard flow-through and centrifuge again 1 min, place the MinElute column in a new 1.5 ml LoBind tube, elute DNA with 18 µl of elution buffer twice, and use 1 µl for DNA concentration measurement with Qubit (*see* **Note 24**).
10. Nick translation: transfer 34 µl of adaptor-ligated DNA into a new 1.5 ml tube, add 4 µl 10× NEBuffer 2 (NEB), 0.8 µl dNTP-Mix (100 mM) and 1 µl DNA Polymerase I (10 U/µl) (NEB), fill with water to 40 µl, mix and incubate reaction for 30 min at 16 °C in a thermomixer; if necessary, inactivate the enzyme at 65 °C for 10 min and store reaction at 4 °C until large-scale PCR setup (*see* Subheading 3.3.3).

3.3.2 Barcoded Preparation of Mate-Paired Libraries

Mate pairs are defined as a pair of DNA fragments that originate from the two ends of the same genomic DNA fragment, the distance between the two mates depends upon the size of the original genomic DNA fragment (insert size), and can range from ~100 bp to several kb. To create mate-paired tags, a linear DNA fragment needs to be circularized by the use of LMP CAP adaptors connected to an internal biotinylated adaptor. These adaptors are provided by the SOLiD Mate-Paired Library Oligo Kit from Applied Biosystems (*see* also Subheading 2.3.2). The resulting DNA circle has one nick in each strand due to the LMP CAP adaptor. During nick translation, the length of the fragments that will be released after T7 exonuclease and S1 nuclease digest depends on the time and temperature of the DNA polymerase reaction. The largest part of the inserted, circularized DNA is cut out enzymatically, which leaves the two ends of the inserted DNA as mate-paired tags. After ligation of universal adaptors and subsequent PCR amplification, the mated tags (2×50 bp) can be sequenced together, allowing computational detection of larger structural genome variants. Each mate-paired library can also be indexed by using modified P2-adaptors containing a unique barcode sequence.

1. Shear 5 µg of extracted fosmid pool DNA in 125 µl nuclease-free water with the HydroShear (standard assembly) to generate gDNA fragment sizes of 1–2 kb (speed code 5, 20 cycles), 3–4 kb (speed code 13, 20 cycles), and 5–10 kb (speed code 15, 5 cycles). Control the shearing results on a 0.8% agarose gel (*see* **Note 25**).

2. Purify sheared fosmid pool DNA with QIAquick Gel Extraction Kit: add 3 volumes QG buffer and 1 volume isopropanol to sheared DNA, transfer 750 µl of DNA-QG buffer mix to QIAquick column (maximum DNA binding capacity 10 µg per column), wait for 2 min at room temperature, centrifuge 1 min at minimally 13,000 × *g* and discard flow-through, repeat the steps until the entire sample has been loaded, wash column with 750 µl PE buffer, centrifuge for 2 min at minimally 10,000 × *g*, repeat centrifugation, air-dry column for 2 min, transfer column to a new 1.5 ml LoBind tube, elute by adding 30 µl EB buffer to column, let stand for 2 min, centrifuge column at minimally 10,000 × *g* for 1 min, repeat elution and centrifugation step.
3. End-repair with End-It DNA End-Repair Kit (Epicentre) as per manufacturer's protocol; the total reaction volume is 10-fold the amount of input DNA (in µg). Purify with QIAquick columns (*see* **step 2**).
4. Ligation of end-repaired sheared fosmid pool DNA to LMP Cap Adaptors resulting in a nick on each DNA strand during circularization: first calculate the molarity of each DNA fragment size (insert) based on the formula $((10^6/660) \times (1/\text{average insert size})) = X$ pmol/µg DNA. Use the molarity to calculate the amount of adaptor needed for the ligation reaction in µl ((total amount of input DNA × pmol of DNA fragment × 100)/(50 pmol) = µl adaptor). Combine 150 µl 2× Quick Ligase Reaction Buffer (NEB), the calculated volume of LMP cap adaptor, 7.5 µl Quick Ligase enzyme (NEB), the total amount of input DNA and add dH_2O to have a 300 µl reaction mix, then incubate the mix at room temperature for 10 min. Purify with QIAquick columns (*see* **step 2**).
5. Size selection: To remove unbound CAP adaptors, prepare a 0.8% TAE agarose gel (Applied Biosystems) with ethidium bromide, mix ligated DNA with 10x loading dye, load 11 µl of sample per well, keep one well empty on each side of the DNA sample, load 1 µl of the 1 kb size DNA Ladder at both sides of the gel, run the gel at 120 V/cm, visualize the gel on an UV illuminator, excise a gel slice corresponding narrowly to the insert size and extract DNA with QIAquick Gel Extraction Kit as per protocol.
6. Circularize LMP CAP adaptor-ligated DNA with a biotinylated internal adaptor: for circularization, the components are calculated per µg of size-selected insert DNA: for a 1–2 kb insert size use 182.5 µl of 2× Quick Ligase Buffer, 1.5 µl Internal Adaptor (2 µM), 9 µl Quick Ligase, and add dH_2O to a volume of 365 µl. Accordingly, circularization of 3–4 kb insert sizes (or 5–10 kb insert sizes, respectively) require 280 µl (360 µl) of 2× Quick Ligase Buffer, 0.65 µl (0.4 µl)

Internal Adaptor, 14 µl (18 µl) Quick Ligase, then add dH_2O to a total volume of 560 µl (720 µl). Incubate the reaction for 10 min at room temperature, purify with QIAquick Gel Extraction Kit (*see* Subheading 3.3.2, **step 2**). The internal adaptor is biotin-labeled, enabling the specific binding of circularized DNA to streptavidin beads later in the protocol (*see* **step 11**) (*see* **Note 26**).

7. Isolate circularized DNA by digesting un-circularized DNA with Plasmid Safe ATP DNAse: combine 5 µl 25 mM ATP, 10 µl 10× Plasmid-Safe Buffer, the total volume of circularized DNA from **step 6**, add 0.33 µl Plasmid-Safe DNAse (10 U/µl) per µg of circularized DNA, then add dH_2O to a total reaction volume of 100 µl and incubate in a thermomixer at 37 °C for 40 min. The DNAse will digest only noncircularized DNA fragments. Purify with QIAquick Gel Extraction Kit Protocol (*see* Subheading 3.3.2, **step 2**) and quantify circularized DNA with Qubit; a minimum of 200 ng circularized DNA should be recovered to proceed with the next steps. For more complex genomes, 600 ng–1 µg circularized DNA is recommended for a high-complexity library.
8. The size of the mate-paired tags to be produced in this step critically depends on the reaction temperature and time of the nick-translation of circularized DNA: work on ice, per 1000 ng of circularized DNA add 5 µl of 100 mM dNTP mix, 50 µl 10× NEBuffer 2, circularized DNA, and bring up to a total volume of 490 µl with dH_2O; chill the reaction for at least 5 min by putting the tube into an ice-water bath, then quickly add 10 µl DNA polymerase I (10 U/µl); mix, incubate the reaction at 0 °C for 12–14 min, immediately stop the reaction by adding 3 volumes of Buffer QG and 1 volume of isopropanol and purify with QIAquick Gel Extraction protocol (*see* Subheading 3.3.2, **step 2**). The size of mate-paired tags created by DNA polymerase I can be controlled by reaction time and temperature (*see* **Note 27**).
9. T7 exonuclease digest to release the mated tags: per 1000 ng of circularized DNA combine 50 µl 10× NE Buffer 4 and circularized DNA, bring up to a total volume of 480 µl with dH_2O, add 20 µl T7 exonuclease (10 U/µl) and mix and incubate the reaction at 37 °C for 30 min. Purify with QIAquick Gel Extraction Kit according to the manufacturer's protocol (*see* Subheading 3.3.2, **step 2**). T7 exonuclease recognizes the nicks within the circularized DNA created in **step 8**, and digests the un-ligated DNA strand away from the tags creating a gap in the sequence. This allows cleavage of the mate-paired tags from the circularized template by S1 Nuclease (*see* the next step).
10. S1 Nuclease digest: Use S1 dilution buffer to prepare S1 Nuclease with an activity of 1 U/µl (Invitrogen); per 1000 ng

circularized DNA mix 50 µl 10× S1 Nuclease buffer, 25 µl 3 M sodium chloride, 50 µl 100 mM magnesium chloride, 20 µl of diluted S1 Nuclease, add dH_2O to a total volume of 500 µl, incubate the reaction at 37 °C for 30 min, immediately stop the reaction by adding 3 volumes of QG Buffer and 1 volume of isopropanol, and proceed with the QIAquick Gel Extraction Kit Protocol (*see* Subheading 3.3.2, **step 2**). S1 Nuclease removes the non-ligated part of the inserted DNA, leaving a linearized molecule with the ends (tags) of the inserted DNA attached to both sides of the molecule (mate-paired tags).

11. Binding of biotinylated library molecules to Streptavidin beads for purification: preparing Streptavidin Dynabeads for use after the end-repair step. Prepare the Streptavidin binding buffer by mixing 10 µl Tris–HCl (500 mM, ph 7.5), 200 µl 5 M sodium chloride, 1 µl 500 mM EDTA and add 289 µl dH_2O to a total volume of 500 µl. Prepare 1× Bead Wash Buffer (from the Solid Buffer Kit) by mixing 5 µl 100× BSA (NEB) and 495 µl dH_2O. Pre-wash Streptavidin Dynabeads: vortex bottle of Streptavidin Dynabeads, transfer 90 µl of beads into a 1.5 ml LoBind tube, add 500 µl 1× Bead Wash Buffer. Vortex and spin down, place tube into magnetic stand (Applied Biosystems) and wait until the solution clears. Aspirate the supernatant and discard it, add 500 µl 1× BSA, and vortex 15 s. Spin down, place tube into magnetic stand until solution clears, aspirate and discard supernatant. Add 500 µl of 1× Bind & Wash Buffer, vortex 15 s, spin down, place tube into magnetic stand to clear solution, aspirate and discard supernatant. Use pre-washed beads in **step 13**.
12. Combine 10 µl End-Repair buffer (10×), 10 µl ATP (10 mM), 10 µl dNTPs (2.5 mM each), 2 µl End-Repair Enzyme Mix, a total volume of S1-digested DNA from step 10, then add dH_2O to a total volume of 100 µl, incubate the reaction mix for 30 min at room temperature, stop the reaction by adding 5 µl EDTA (500 mM), 200 µl Streptavidin binding buffer (prepared in **step 11**) and 95 µl dH_2O.
13. Binding the library DNA molecules to Streptavidin beads: add the entire 400 µl reaction from **step 12** to the pre-washed Streptavidin beads (prepared in **step 11**), vortex, place the tube into a rotator for 30 min at room temperature, and spin down the tube.
14. Wash the DNA-Strepatavidin bead complex: place tube with DNA bound to beads into magnetic stand until solution clears. Remove and discard supernatant with pipet, remove tube from magnetic rack, and add 500 µl 1× Bead Wash Buffer (prepared in **step 11**). Transfer the suspension to a new 1.5 ml LoBind tube, vortex for 15 s, spin down, and resuspend in 500 µl 1× Bind & Wash Buffer. Vortex and spin down the solution and

clear on magnetic rack. Discard the supernatant again, and repeat washing step with another 500 µl 1× Bind & Wash Buffer. Resuspend the beads in 500 µl 1× Quick Ligase Buffer, vortex for 15 s, spin down and clear suspension on magnetic rack. Discard supernatant a last time, and resuspend beads in 97.5 µl 1× Quick Ligase Buffer.

15. Ligate P1 and P2 adaptors to end-repaired mate-paired molecules: calculate the necessary volume of P1 and P2 adaptors by using the formula in Subheading 3.3.1, **step 6**, with the size of circularized DNA corresponding to the average insert size. Combine 97.5 µl of the DNA-bead complex, the calculated volume of P1 and P2 Adaptor (50 pmol/µl each) and 2.5 µl Quick Ligase, and incubate the reaction for 15 min at room temperature (*see* **Note 28**).
16. Wash the DNA-bead complex as described in **step 14** but modified in the last step: add 500 µl 1× NEBuffer 2 (instead of Quick Ligase Buffer) and resuspend beads in 96 µl 1× NEBuffer 2.
17. Nick translation of library: Add 2 µl 100 mM dNTP mix and 2 µl DNA Polymerase I (10 U/µl) to the 96 µl DNA-bead complex in NEBuffer, incubate the reaction for 30 min at 16 °C, place tube into magnetic stand until solution clears, remove and discard supernatant, resuspend beads in 500 µl Buffer EB, place tube into magnetic stand until cleared, remove supernatant, and resuspend beads in 30 µl Buffer EB.

3.3.3 Large-Scale PCR of Fragment and Mate-Paired Libraries

The library is amplified with the Invitrogen PCR SuperMix. The number of PCR cycles should be as small as possible to avoid PCR-related biases due to differential amplification of library molecules.

Fragment Libraries:

1. To prepare a PCR master mix, pipet 400 µl PCR SuperMix into a 1.5 ml tube, add 10 µl of each primer, 40 µl dH_2O and 2.5 µl Pfu Polymerase (2.5 U/µl). Before adding the sample, transfer 50 µl of master mix without DNA sample into a new 200 µl tube (negative PCR control). Add 40 µl of nick-translated DNA (*see* Subheading 3.3.1, **step 10**) to the remaining master mix, vortex, and transfer four aliquots into new 200 µl PCR strips (*see* **Note 29**). Run large-scale PCR program (5 min at 95 °C, hold; few cycles (*see* **Note 30**) 15 s at 95 °C, 15 s at 62 °C, 60 s at 70 °C; 5 min at 70 °C, hold; 4 °C, forever, hold); the number of PCR cycles critically depends on the amount of input DNA as measured after purification of nick-translated DNA.
2. Purify PCR reaction with MinElute columns, elute in 60 µl, add 10 µl loading dye (6×), load the sample on a 4% Reliant NuSieve 3:1 agarose gel, run for 40 min at 110 V, size-select

PCR product between 150 and 200 bp and control for over-amplification; do not proceed, if PCR generates an over-amplification band. Gelase digest and extract DNA sample as described (*see* Subheading 3.3.3, **step 4**).

Mate-paired libraries:

1. Prepare a PCR master mix for four PCR reactions: mix 200 µl PCR SuperMix (Invitrogen), 8 µl of Library PCR Primer 1 (50 µM), 8 µl Library PCR Primer 2 (50 µM), 1 µl cloned Pfu DNA polymerase (2.5 U/µl), and 143 µl dH_2O. Aliquot 90 µl of PCR master mix to a 200 µl PCR tube, add 10 µl dH_2O as negative control. To the remaining 270 µl PCR master mix, add 27 µl DNA-bead complex, vortex, pipet 90 µl aliquots into three PCR tubes, run PCR program with initial denaturing step (10 min at 95 °C) followed by 5 cycles (15 s at 95 °C, 15 s at 62 °C, 60 s at 70 °C); 4 min at 60 °C, hold; 4 °C, forever, hold (*see* **Note 30**).
2. Control PCR reaction: Use a 4 µl aliquot of PCR reaction and mix with 1 µl 6× Orange DNA Loading Dye. Load sample on a 2.2% FlashGel System, load Low Range DNA Ladder in adjacent well, and run the gel for 6 min at 275 V. If fairly robust PCR amplification products are visible on the gel, pool all three PCR reactions into a new 1.5 ml LoBind tube. Otherwise, run two additional cycles on thermal cycler, control an aliquot using a Lonza FlashGel System, repeat thermal cycling steps until amplification is observed.
3. Place 1.5 ml tube on magnetic rack, carefully aspirate supernatant and transfer into a new 2.0 ml LoBind tube. Purify with QIAquick Gel Extraction protocol (*see* Subheading 3.3.2, **step 2**).
4. Gel purification step: use a Reliant 4% NuSieve 3:1 agarose gel with ethidium bromide, mix 6× Orange DNA Loading Dye to mate-paired library to a final 1× concentration. Load the entire dye-mixed sample DNA in aliquots of 11 µl per gel slot. Keeping one well on each side of sample DNA empty, load 2 µl Low Range DNA Ladder. Run the gel at 120 V, visualize and excise the sample DNA between bands 275 and 300 bp. Weigh gel slices in tared 15 ml conical tubes, add 6 volumes QG buffer per 1 volume gel, and dissolve the gel slices by vortexing at room temperature (do not heat). Add 1 gel volume isopropanol, and mix and invert. Apply 750 µl to a QIAquick column, let stand for 2 min at room temperature, centrifuge for 1 min at minimally 10,000 × *g*. Repeat until the entire sample has been loaded on column. Wash the column by adding 500 µl QG buffer, centrifuge at 10,000 × *g*, wash with 750 µl PE buffer, centrifuge, and air-dry the column for 2 min. Place column on a new 1.5 ml LoBind tube, add 30 µl EB buffer, wait 2 min, and centrifuge for 1 min. Quantify the eluted library with the Qubit system as per manufacturer's protocol.

3.3.4 Preparation of Fragment and Mate-Paired Sequencing Libraries for Emulsion PCR

After large-scale PCR purification, library quantification by the use of the Qubit system is sufficiently accurate to calculate the dilution required to obtain picogram concentrations of the library. Alternatively, TaqMan or SybrGreen Real Time PCR can be applied for quantification.

To prepare 500 pM fosmid DNA input samples, measure amplification products by real-time PCR or Qubit fluorometry as per manufacturer's protocol. Prepare several aliquots of 500 pM dilutions of amplified DNA as follows: dilute necessary volume of sample DNA in a volume of 30 µl low TE buffer to yield a final concentration of 5 ng/µl. Pipet 8 µl of the prepared dilution into a new 100 µl tube, and add 32 µl low TE buffer to prepare a 1 ng/µl dilution. Pipet 4.8 µl of 1 ng/µl dilution into a new tube, and add 75.2 µl low TE buffer to a final concentration of 60 pg/µl (500 pM for fragment libraries) (*see* **Note 31**).

3.4 Processing Next Generation Sequencing Libraries for Instrument Run

The SOLiD technology relies on sequencing library amplification on a solid support. To this end, an aqueous and an oil phase are mixed to form an emulsion. During emulsion PCR (ePCR), ideally, multiple copies of one single DNA template molecule are generated on one magnetic bead contained in one emulsion droplet. To avoid multi-clonal amplification, precise DNA quantification and input amounts are essential requirements. Un-templated beads are separated and discarded in a subsequent enrichment step. To ensure that the beads will deposit on the sequencing slide, a 3′modification of the enriched templated beads is performed. For control, a workflow analysis (WFA) run is carried out with a small portion of the templated beads to analyze the templated bead quality (e.g., multi-clonal beads result in poor sequencing outcome), and the bead concentration.

3.4.1 Emulsion PCR

1. Prepare oil phase (all components provided by the SOLiD ePCR Kit): pipet 35 ml of oil into a 50 ml Falcon tube, add 1.8 ml emulsion stabilizer 1 by using a 2.5 ml glass pipet, and add 400 µl of stabilizer 2 with a 1 ml glass pipet. Avoid air bubbles when aspirating stabilizer; in case bubbles have been drawn into pipet, dispense and re-aspirate. Fill Falcon tube with oil to 40 ml, close tube and vortex vigorously to emulsify oil components; open cap and let the Falcon tube degas for at least 30 min. Pipet 9 ml of oil phase with a 10 ml syringe into a SOLiD ePCR tube and cap, place tube on IKA Turrax.
2. During degasing of the oil phase, prepare aqueous phase containing the fosmid pool DNA sample. Mix reaction components depending on sequencing library molarity with a total reaction volume of 2720 µl. For a 1 pM sequencing library combine 280 µl 10× PCR Buffer, 392 µl 100 mM dNTP mix (25 mM per dNTP), 70 µl 1 M MgCl, 16.8 µl ePCR Primer 2 (500 µM), 11.2 µl ePCR Primer 1 dilution (10 µM), 5.6 µl

sequencing library template molecules (500 pM), 1,644.4 µl nuclease-free water and 300 µl AmpliTaq Gold (5 U/l) in a 15 ml Falcon tube, mix by gently inverting the closed tube.

3. Prepare P1 beads: vortex one vial of SOLiD P1 beads, spin down tube, place tube in magnetic rack, wait 1 min until the solution has cleared, discard supernatant, resuspend the beads in 200 µl 1× Bead Block Solution. Vortex, spin down, and place tube in Covaris S2 Sonicator. Run the Bead Block Declump program, place tube in magnetic rack, and wait until solution clears. Discard supernatant, resuspend in 200 µl 1× TEX buffer, vortex, and spin down.
4. Sonicate the prepared P1 beads with the Bead Declump program, immediately add 80 µl of sonicated P1 beads to aqueous phase (*see* **step 2**). Mix gently by swirling the reaction, and use an Eppendorf semiautomated Xstream pipettor with a 10 ml tip to aspirate 2.8 ml of aqueous phase with bead-DNA complex and P1 beads. Start the IKA Turrax to swirl the oil phase, make sure the program runs at least 5 min. When the IKA Turrax reaches full spinning speed, place Xstream pipettor tip in the middle of ePCR tube and dispense as programmed. Let the oil-aqueous phase mix swirl until IKA Turrax stops (after 5 min).
5. Use a 5 ml tip on Eppendorf Repeater Plus Pipette, dispense 100 µl emulsion PCR mix per well into a 96-well PCR reaction plate, check bottom of the 96-well plates for air bubbles, spin down if necessary, and run emulsion PCR in a thermal cycler with gold/silver block (program: 5 min at 95 °C, hold; 40 cycles 15 s at 93 °C, 30 s at 62 °C, 75 s at 72 °C; 7 min at 72 °C, hold; 4 °C, forever, hold).
6. After ePCR finishes, check wells for broken emulsion reactions. Instead of a homogeneous milky suspension, a broken emulsion is indicated by three different layers in a well: a milky suspension, followed by a clear liquid layer (an aqueous phase) and dark freckles from the beads at the bottom of the well. Remove the content of "broken wells" before the next step; do not proceed to the next step if more than four wells show broken emulsion reactions.

3.4.2 Breaking the Emulsion PCR

1. Under a fume hood, pipet 100 µl 2-butanol into each well of the 96-well ePCR plate. Pipet up and down to mix, then pipet the broken emulsion PCR suspensions into a 50 ml tray. Transfer into a new 50 ml Falcon tube, and rinse tray with additional 2-butanol to collect remaining beads from tray. Fill the Falcon tube to 30 ml with 2-butanol, cap and vortex, centrifuge for 5 min at 2000 × *g*, and carefully decant the supernatant (oil). Turn Falcon tube and place for 5 min onto a paper towel to drain remaining oil.

Alternatively use Emulsion Collection Tray: place blue metal adaptor on the ePCR 96-well plate, then place the Emulsion Collection Tray like a cap (bottom up) on the top of the ePCR 96-well plate with metal adaptor. Use parafilm around metal adaptor to seal the connected plates, and flip the plates so that the ePCR plate is upside-down over the Emulsion Collection Tray. Centrifuge the plate construction for 2 min at 550×*g* (centrifuge with 96-well plate adaptor), place the assembly under a fume hood, remove the ePCR plate, and add 10 ml of 2-butanol to the Collection Tray containing centrifuged ePCR emulsion. Pipette the mixture up and down until it appears homogeneous, transfer to a 50 ml Falcon tube, and rinse reservoir with 2-butanol to collect remaining beads. Cap and vortex Falcon tube, centrifuge at 2000×*g* for 5 min, decant the oil phase and place inverted tube on paper towel, and then wait 5–10 min.

2. Wash templated beads: Resuspend the beads in 600 µl 1× Bead Wash Buffer, carefully pipet up and down and transfer beads to a fresh 1.5 ml LoBind tube, rinse the remainder at the bottom of the 50 ml Falcon tube with additional 600 µl 1× Bead Wash Buffer and transfer to 1.5 ml LoBind tube, vortex, centrifuge at 21,000×*g* for 1 min. Remove the oil phase with pipet, change pipet tip, discarding supernatant. Repeat bead washing as described but use 150 µl to resuspend and transfer the beads, and 150 µl to transfer remaining beads, add 1 ml 1× Bead Wash Buffer, centrifuge at 21,000×*g* for 1 min, discard supernatant with pipet, resuspend in 200 µl 1× TEX buffer, put tube on magnetic rack until it clears, discard supernatant, add 200 µl 1× TEX buffer and sonicate the beads with program Declump 1. Determine the bead concentration using a 1:10 dilution on an UV-spectrophotometer (NanoDrop), and compare the color of the bead dilution to photographed colors for different bead concentrations. Optimally, the concentration reflects closely the input amount of beads, about 1.6 billion, indicated by a "medium" color. Adjust the bead concentration by adding 1× TEX in case the color is too dark, or by placing tube in magnetic rack and removing small volume of supernatant in case the color is too light, to match the bead suspension to required volume. Quantify with NanoDrop.

3.4.3 Enrichment of Templated Beads

1. Prepare buffers and enrichment beads (from the SOLiD Bead Enrichment Kit) for use in subsequent steps: for a single ePCR reaction, transfer 1.8 ml denaturing buffer into a 15 ml Falcon tube, add 200 µl denaturant, cap and vortex denaturing reagent. Prepare a fresh 60% glycerol solution, transfer 4 ml dH_2O into a 15 ml Falcon tube and add 3 ml glycerol twice, cap and vortex. For preparation of the enrichment beads, vortex the beads and immediately transfer 300 µl of the beads to

a new, 1.5 ml LoBind tube, centrifuge for 5 min at 21,000 × *g*, discard the supernatant, resuspend in 900 μl 1× Bind and Wash Buffer, centrifuge for 5 min at 21,000 × *g*, repeat Bind and Wash step, resuspend in 150 μl 1× Bind and Wash Buffer, add 1.5 μl 1 mM Enrichment Oligo, vortex and place on a rotator at room temperature for 30 min. Centrifuge beads for 5 min at 21,000 × *g*, remove supernatant, resuspend in 900 μl 1× TEX buffer, repeat centrifugation and TEX wash step, and resuspend beads in 75 μl 1× Low Salt Binding Buffer.

2. Prepare templated beads for enrichment step: place tube with templated beads in magnetic rack, remove supernatant, resuspend in 300 μl freshly prepared denaturing buffer, wait 1 min, place tube in magnetic rack, remove supernatant, repeat resuspension and magnetic rack step, resuspend templated beads in 300 μl 1× TEX buffer, place in magnetic rack and remove supernatant, repeat resuspension in TEX and magnetic rack step. Resuspend the templated beads in 150 μl TEX and transfer to a new 0.5 ml LoBind tube. Declump beads with Program 1 on Covaris.

3. Pipet enrichment beads (from **step 1**) into 0.5 ml LoBind tube containing templated beads, vortex, spin down and sonicate the mixture using the Covalent Declump 3 program, pulse-spin and incubate at 61 °C, vortex and spin down the mixture every 5 min over a time period of 15 min. Cool beads on ice for 2 min. Use a new 1.5 ml LoBind tube and add 400 μl of the 60 % glycerol solution; do not vortex. Mix cooled beads by pipetting up and down, carefully load the total volume on the top of the glycerol-filled tube. Do not vortex at this point, but centrifuge for 3 min at 21,000 × *g*. Prepare a 2.0 ml LoBind tube with 1 ml 1× TEX buffer, and transfer top layer of beads (swimming on glycerol) to the bottom of tube prepared with TEX (be cautious to not transfer the un-templated beads sitting at the bottom of the glycerol tube). Fill up to 2.0 ml with 1× TEX, vortex and centrifuge transferred templated beads at 21,000 × *g* for 1 min. If the beads are not sufficiently pelleted (in case of a carry-over of too much glycerol), divide them into two halves and fill each one into a tube, add 500 μl 1× TEX buffer to each tube, vortex and centrifuge at 21,000 × *g* for 1 min. Discard supernatant, and add 200 μl 1× TEX buffer to each tube. Resuspend the beads and pool the two halves again into one single tube. Otherwise, if the beads are pelleted, directly remove supernatant and add 400 μl 1× TEX buffer. Proceed with protocol as follows: vortex and centrifuge tube at 21,000 × *g* for 1 min, remove supernatant, resuspend with 400 μl Denaturing Buffer (which is prepared fresh by mixing 1.8 ml Denaturation Buffer with 200 μl Denaturant and vortexing), let stand for 1 min, place

in magnetic rack, remove supernatant, repeat denaturing and magnetic clarification until all white enrichment beads are removed. Resuspend in 400 µl 1× TEX, place in magnetic rack, discard supernatant, repeat resuspension in TEX and magnetic rack step, resuspend templated beads in 200 µl 1× TEX, vortex, spin down and transfer into new 1.5 ml LoBind tube. Declump beads using program Declump 1, place tube in magnetic rack, remove supernatant and resuspend in 400 µl 1× TEX, place in magnetic rack, remove supernatant. Repeat the last two steps until supernatant is clear, and finally resuspend beads in 400 µl 1× TEX (*see* **Note 32**).

3.4.4 3′-End Modification of Enriched Templated Beads

The templates on the selected beads are subjected to a 3′-end modification to allow covalent bonding to the slide.

Sonicate enriched templated beads (*see* Subheading 3.4.3) with Declump 3, pulse-spin and prepare 500 µl 1× Terminal Transferase Reaction (TTR) mix (from the SOLiD Bead Deposition Kit) per ePCR reaction. Mix 55 µl 10× TTR, 55 µl 10× cobalt chloride, and 390 µl dH_2O to a total volume of 500 µl. Prepare 1 mM Bead Linker solution by mixing 1 µl Bead Linker (50 mM) to 49 µl low TE buffer. Place the tube with enriched beads in magnetic rack, remove supernatant, resuspend in 100 µl 1× TTR buffer, and transfer reaction to a new 1.5 ml LoBind tube. Place in magnetic rack, remove supernatant, repeat 1× TTR resuspension and magnetic rack step, resuspend beads in 178 µl 1× TTR, and add 20 µl Bead Linker solution (1 mM). Sonicate the mixture with Declump 3, add 2 µl Terminal Transferase enzyme (20 U/µl). Vortex, pulse-spin, rotate tube for 2 h at 37 °C (place rotator in incubator), then place tube in magnetic rack, discard supernatant, resuspend in 400 µl 1× TEX, perform magnetic clearing, resuspend in 200 µl 1× TEX. Quantify beads after utilizing program Declump 1 on NanoDrop and compare with SOLiD bead color chart.

3.4.5 Bead Deposition on SOLiD Sequencing Slide and Instrument Run

1. Declump enriched templated beads with program Declump 1, pulse-spin, transfer appropriate volume of beads (as determined after NanoDrop quantification or in a WFA run) to a new 1.5 ml LoBind tube, place in magnetic rack, remove supernatant, resuspend beads in 400 µl Deposition Buffer, vortex well, spin down, place in magnetic rack, discard supernatant. Repeat resuspension in Deposition Buffer and magnetic rack step twice according to the number of "fields" on the sequencing slide to be filled with beads; the beads need to be resuspended in different volumes, 300 µl per well in a 8-well deposition chamber, 400 µl per well in a 4-well, and 550 µl per well in a single-well chamber.
2. Insert a sequencing slide into the slide carrier and place the assembled slide carrier into a deposition chamber base, and on top place the appropriate deposition chamber lid

(1-well, 4-well, and 8-well). Sonicate beads with program Declump 3, pulse-spin and sonicate again with Declump 3, pipette the bead solution up and down, immediately fill bead solution into corresponding well of deposition chamber to avoid clumped beads on slide, seal portholes and incubate at 37 °C for 1.5 h.

3. In the meantime, prepare the SOLiD sequencing instrument and set up SOLiD sequencing run as per manufacturer's protocol. After the incubation step has finished, remove adhesive seals, drain the top of the deposition chamber with Deposition Buffer, use a 1000 ml pipette, press on portholes, and aspirate the entire Deposition Buffer from deposition wells until freshly layered Deposition Buffer is drawn into the wells, loosen lid of deposition chamber, immediately place the slide carrier assembly on the flow cell of the SOLiD instrument. Avoid drying out the slide, close flow cell, and proceed with the specific steps of the instrument run as described in the manufacturer's protocol.

3.5 Computational analysis of Fosmid Sequences and Haplotype Assembly

The bioinformatics procedure starts from the raw reads obtained from barcoded high-throughput sequencing of each fosmid pool. The main steps include identification of heterozygous sites, fosmid detection, calling of fosmid-specific genotypes at heterozygous sites, and haplotype assembly using the predicted fosmids and their allele calls as virtual reads. Detailed steps to perform this procedure are summarized in Fig. 6 and described as follows:

3.5.1 De-Indexing Barcoded Fosmid Pools

Separate sequence reads per fosmid pool, identifying the barcode sequences in the first 5 bp of each read. Reads sharing the same barcode belong to the same fosmid pool. For SOLiD reads, this is done using Bioscope 1.3. For Illumina reads, this can be done with the "Deconvolute" command of NGSEP.

3.5.2 NGS Read Alignment

Align the reads per pool against the reference genome. For SOLiD reads, use Bioscope 1.3. For reads sequenced using other platforms such as Illumina, align the reads using software tools for alignment of standard WGS reads such as bwa [17] or bowtie2 [14]. Then, sort the alignments by reference sequence and position using the sort command of samtools or the Picard software. In both platforms the final output of this step is a file in SAM or BAM format for each pool, containing the information of the reads aligned to the reference genome (*see* **Note 33**).

3.5.3 Identification of Heterozygous Variants

Merge the BAM files obtained for each pool using the merge command of samtools or Picard. Then, identify variants against the reference genome, again using Bioscope 1.3 for SOLiD reads. Alternatively (*see* **Note 34**), variants can be identified from aligned standard WGS read data by the use of the NGSEP pipeline [15], samtools [17], or the GATK pipeline [16] for other platforms such

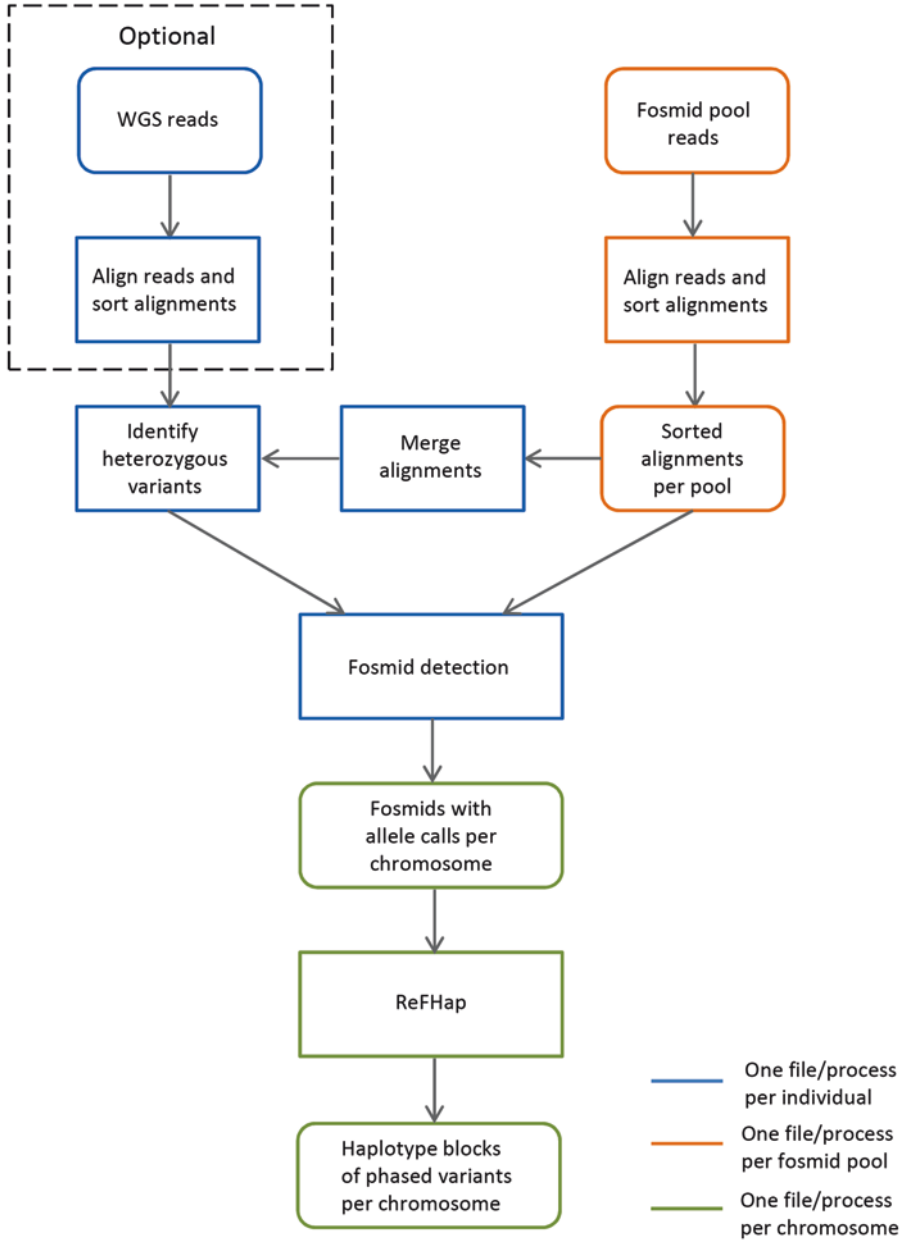

Fig. 6 Overview of computational steps involved in analysis of fosmid pool-based sequences and phasing. Complementary analyses of whole genome sequencing (WGS) data are presented as an option

as Illumina. Detailed instructions, scripts, and recommended parameters for command line usage of NGSEP can be found at (http://sourceforge.net/projects/ngsep/files/training/). Regardless of the sequencing platform and analysis pipeline used at this stage, the final output is a list of variants in VCF format and one sorted BAM file for each fosmid pool (*see* **Note 35**).

3.5.4 Fosmid Detection

Run the fosmid detection program following the specific instructions in the README file. This program receives the list of variants against the reference genome and an XML file describing, for each sequenced fosmid pool, the identification number (id), average coverage, and a path to the BAM file containing the aligned and sorted reads. A GFF file with the format produced by Bioscope 1.3 including predicted heterozygous deletions for the individual can be added as a parameter. If, moreover, files with allele calls in the same format are registered for each pool at the XML file, the fosmid detector can include these large deletions in the final haplotypes. With all this information, the fosmid detector performs the tasks described in detail earlier [3]. The following internal steps are being performed by the fosmid detector and summarized here:

(a) Extract the heterozygous variants from the VCF file obtained in the previous section.

(b) Predict the physical locations of the fosmids sequenced in each pool from the aligned reads of the pool and call one (usually homozygous) fosmid-specific genotype (FSG) for each heterozygous variant covered by one predicted fosmid. To predict fosmids, the software divides the genome in nonoverlapping segments of a fixed length (1 kb by default) and processes the alignments to calculate the read coverage of each segment. Then, it tags bins as candidate members for membership to a fosmid if the read coverage exceeds a minimum threshold (0.4 by default). Fosmids are called collecting nearly adjacent bins (with a maximum gap of 9 bins by default) extending between a minimum and a maximum threshold (by default 3–60 kb). Stretches that are longer than the maximum length are split into two or more predicted fosmids with lengths within the minimum and maximum thresholds. FSGs can be called either with NGSEP or with the allele coverage procedure described [3] (*see* **Note 35**). Although most FSGs should be homozygous, heterozygous FSGs are possible if two fosmids from the two parental homologues of a chromosome overlap within one pool. Consequently, bins showing heterozygous calls within a pool are tagged, and the tagged fosmids split into two fosmids, the first predicted from the covered bins extending upstream of the first tagged bin, and the second from the covered bins following the last tagged bin. This step is important to avoid switch errors produced by overlaps of complementary fosmids. After these steps, fosmids containing only one homozygous FSG are removed to obtain the set that will be used for single individual haplotyping.

(c) Finally, the fosmid detector allows generation of the input matrix for phasing separately for each chromosome, with the fosmids (one per row) sorted by the position of the first genotyped variant, and the allele calls (one per column) designated

by physical genome position. This procedure results in the production of the variation matrixes that serve as input files for haplotype assembly by use of ReFHap.

3.5.5 Phasing by Use of ReFHap

For each chromosome, take the input file for ReFHap generated by the fosmid detector as described above (4c) and run ReFHap following the instructions available in the README file. Details of the algorithms implemented in ReFHap are available in Duitama et al., 2010 [11], and benchmarks with other tools/SIH algorithms can be found in Duitama et al., 2011 [5]. For each chromosome, ReFHap outputs the blocks of variants that could be phased by tiling the fosmids (*see* Fig. 7a), and for each block, it outputs the chromosomal position of each phased variant, followed by the allele in the first haplotype, and the allele in the second haplotype (*see* Fig. 7b).

a

chr1	FosSeq	ctg	1	40999	1	+	id=chr1_BC26_6958:posn=1-40999:reads=970:bins=40:cov=0.99:poolId=BC26
chr1	FosSeq	ctg	1	30999	1	+	id=chr1_BC28_10738:posn=1-30999:reads=482:bins=30:cov=0.65:poolId=BC28
chr1	FosSeq	ctg	1	52999	1	+	id=chr1_BC29_12761:posn=1-52999:reads=3026:bins=52:cov=2.41:poolId=BC29
chr1	FosSeq	ctg	1	54999	1	+	id=chr1_B17_37073:posn=1-54999:reads=1260:bins=54:cov=1.09:poolId=B17
chr1	FosSeq	ctg	1	30999	1	+	id=chr1_B22_44150:posn=1-30999:reads=685:bins=30:cov=1.05:poolId=B22
chr1	FosSeq	ctg	1	59999	1	+	id=chr1_B23_45772:posn=1-59999:reads=1376:bins=59:cov=1.1:poolId=B23
chr1	FosSeq	ctg	1	25999	1	+	id=chr1_B27_52229:posn=1-25999:reads=423:bins=25:cov=0.77:poolId=B27
chr1	FosSeq	ctg	1	55999	1	+	id=chr1_B28_53864:posn=1-55999:reads=1915:bins=55:cov=1.64:poolId=B28
chr1	FosSeq	ctg	1	38999	1	+	id=chr1_B30_57409:posn=1-38999:reads=1843:bins=38:cov=2.26:poolId=B30
chr1	FosSeq	ctg	1	45999	1	+	id=chr1_B35_65046:posn=1-45999:reads=1084:bins=45:cov=1.1:poolId=B35
chr1	FosSeq	ctg	1	29999	1	+	id=chr1_B37_68417:posn=1-29999:reads=995:bins=29:cov=1.54:poolId=B37
chr1	FosSeq	ctg	1	47999	1	+	id=chr1_B39_71786:posn=1-47999:reads=1302:bins=47:cov=1.27:poolId=B39
chr1	FosSeq	ctg	1	39999	1	+	id=chr1_B40_73387:posn=1-39999:reads=1677:bins=39:cov=1.96:poolId=B40
chr1	FosSeq	ctg	1000	42999	1	+	id=chr1_BC22_1785:posn=1000-42999:reads=635:bins=42:cov=0.63:poolId=BC22
chr1	FosSeq	ctg	1000	38999	1	+	id=chr1_B13_31958:posn=1000-38999:reads=521:bins=38:cov=0.66:poolId=B13
chr1	FosSeq	ctg	1000	28999	1	+	id=chr1_B19_40285:posn=1000-28999:reads=528:bins=28:cov=0.89:poolId=B19
chr1	FosSeq	ctg	1000	29999	1	+	id=chr1_B29_55605:posn=1000-29999:reads=576:bins=29:cov=0.96:poolId=B29
chr1	FosSeq	ctg	1000	27999	1	+	id=chr1_B33_61712:posn=1000-27999:reads=345:bins=27:cov=0.59:poolId=B33
chr1	FosSeq	ctg	1000	53999	1	+	id=chr1_B36_66810:posn=1000-53999:reads=1522:bins=53:cov=1.34:poolId=B36
chr1	FosSeq	ctg	1000	10999	1	+	id=chr1_B38_70131:posn=1000-10999:reads=139:bins=10:cov=0.64:poolId=B38
chr1	FosSeq	ctg	2000	28999	1	+	id=chr1_BC30_15089:posn=2000-28999:reads=324:bins=27:cov=0.5:poolId=BC30
chr1	FosSeq	ctg	4000	22999	1	+	id=chr1_BC27_8937:posn=4000-22999:reads=132:bins=19:cov=0.29:poolId=BC27
chr1	FosSeq	ctg	4000	29999	1	+	id=chr1_B25_48983:posn=4000-29999:reads=205:bins=26:cov=0.36:poolId=B25
chr1	FosSeq	ctg	6000	56999	1	+	id=chr1_B14_33493:posn=6000-56999:reads=983:bins=51:cov=0.93:poolId=B14

b

BLOCK: offset: 2 len: 2 phased: 2		
536602	0	1
556655	1	0
********		
BLOCK: offset: 4 len: 231 phased:225		
715887	1	0
716595	1	0
724429	1	0
747799	0	1
747911	1	0
748300	1	0
748331	1	0
766409	0	1
779714	1	0
779737	1	0
780559	1	0
786902	1	0
804531	1	0
805284	0	1
805386	1	0
805419	1	0
805558	1	0
805591	0	1
805892	1	0
805897	1	0
805898	1	0
805905	0	1
805897	1	0
...		

Fig. 7 Examples of phasing output files. (**a**) Output file fosmid contigs. A part of a typical output file is shown, listing per chromosome the detected contigs as an output of the fosmid detection program (*see* Subheading 3.5.4). The chromosomal start and end positions per contig are given in columns 4 and 5 by indicating the specific bins (first and last) as intervals of fixed length (1000 bp) that are covered by the fosmid contig. In the following columns, additional properties such as number of reads, bins, read coverage and fosmid super-pool identification number are shown. (**b**) Output file ReFHap with phased variants. A part of a typical output file is shown, listing the haplotype blocks per chromosome with the numbers of phased variants per block. In the first column from the *left*, the chromosomal position numbers of heterozygous variants are given, and in the adjacent, second and third columns the two haplotypes, with the variant alleles (different from the reference sequence) denoted by "1" and the reference alleles denoted by "0." Thus, the second column shows the specific combinations of variants from top to bottom for "Haplotype 1" and the third column the specific combinations of variants constituting "Haplotype 2"

4 Notes

1. Optimally, one kit would be sufficient to generate ten individual fosmid libraries. In practice, the kit will allow generating three to four libraries.
2. Do not vortex the sample to avoid shearing the HMW gDNA.
3. Do not transfer any flocculent material.
4. We find that the gDNA cannot be resuspended in buffer, if the alcohol is not completely evaporated.
5. Alternatively, the Tube Rotisserie Rotator (VWR) can be used, which is also employed during SOLiD sequencing library preparation.
6. Alternatively run a large 1% agarose gel of at least 20 cm length to allow for sufficient resolution, at 30 V/cm overnight.
7. Since under conditions of manual shearing, i.e., variable pressure and speed, the shearing results will turn out to a certain extent variable, check the results on an agarose gel of at least 20 cm length overnight. In order to standardize our manual shearing procedure (which we used at the outset), we constructed an automated prototype, which is not generally available. Therefore, collecting information from experts in the field [18, 19], we have chosen to refer to commercially available HydroShear devices here and the settings that have been experimentally verified.
8. Avoid unnecessary pipetting or vortexing after end-repair, pipet carefully and gently to avoid mechanical damage.
9. If no preparative comb is available, tape 2 wells of a 1.5 mm thick 20-well comb together, keep one well to each side of sample empty, first and last well are used for size ladder. The slots of the preparative comb have a volume of about 200 μl.
10. The gel is very fragile, therefore, do not lift the gel, but transfer it on the gel tray of the electrophoresis chamber.
11. Sheared gDNA appears as a smear; try to cut out gel pieces as exactly as possible to minimize subsequent gelase digestion. Do not stain the gel with ethidium bromide and avoid UVB illumination, which will damage the DNA and significantly reduce ligation efficiency.
12. Scrape glycerol stock of EPI100 bacteria with heat sterilized inoculating loop only briefly before streaking over agar plate to ensure growth of single colonies.
13. Prepare Epi100 bacteria for mass infection 2 days before advancing to the phage packaging step.

14. In order to evaporate condensed water due to storage, place the agar plates in a laminar flow bench one hour before use, remove the cover plates, and let them dry.
15. It is recommendable to produce a total of 1.5×10^6 phages to have a bit of excess later during the fosmid clone partitioning step.
16. The amplification of fosmid clones to high numbers in liquid culture can introduce preferential amplification, that is, overgrowth of few fosmid clones in a fosmid pool, and may lead to significant library complexity reduction. Our experimental results have shown, however, that this was only the case if the second amplification step, required to generate sufficient copies of the ~15,000 fosmid clones in a super-pool before isolation of fosmid DNA for NGS, was also made in liquid cultures. In that case, library complexity was substantially reduced. If the second required amplification step was, however, performed using agar plates, both haploid genomes were found to be nearly completely represented, with an equal physical coverage of both haplotypes. Specifically, we have performed high-throughput SNP typing of the 4 Mb MHC region in one 96-well plate of the fosmid pool-formatted library, demonstrating presence of MHC fosmid clones according to expectations in the fosmid pools and nearly complete and equal coverage of both MHC haplotypes.
17. The wet-lab-based removal of Epi 100 genomic DNA is only partially sufficient; the remaining Epi100 genome, as evident in NGS reads, needs to be removed computationally.
18. Make sure to check whether the tubes are suitable to withstand the centrifugal forces.
19. Use 5 l bottles with cooled, distilled water stored at 4 °C to be able to start the Covaris system more quickly.
20. To prepare a 4 % agarose gel, mix buffer and agarose, let ingredients stand overnight, heat the agarose carefully in microwave, avoid over-boil by checking every 2 s.
21. The E-Gel Size Selection Gel or Flash Gel Recovery System can be used instead; in our experience the recovery was not superior to the gel-based size selection.
22. We created 48 different barcode tags within the P1-Adaptor sequence (which we shortened) before SOLiD barcoded adaptors became commercially available. Thus, 48 pools/sequencing libraries could be parallel processed, and up to 16 barcoded NGS libraries could be multiplexed later in a single sequencing run. To avoid color imbalances on the SOLiD NGS system, all 16 barcoded adaptors should also be employed (per run) in the case where the number of samples is smaller.

Barcode ID		bp	Barcode ID		bp
5′	3′				
P1-BC-1-5′Oligo	AAGAGGATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-25-5′Oligo	TGATGAATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-1-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCCTCTT	29	P1-BC-25-3′Oligo	CCTCTCTATGGGCAGTCGGTGATTCATCA	29
P1-BC-2-5′Oligo	AGTGGTATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-26-5′Oligo	AGCCCGATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-2-3′Oligo	CCTCTCTATGGGCAGTCGGTGATACCACT	29	P1-BC-26-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCGGGCT	29
P1-BC-3-5′Oligo	GTTATAATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-27-5′Oligo	AAACTTATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-3-3′Oligo	CCTCTCTATGGGCAGTCGGTGATTATAAC	29	P1-BC-27-3′Oligo	CCTCTCTATGGGCAGTCGGTGATAAGTTT	29
P1-BC-4-5′Oligo	GCGGTCATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-28-5′Oligo	TACTACATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-4-3′Oligo	CCTCTCTATGGGCAGTCGGTGATGACCGC	29	P1-BC-28-3′Oligo	CCTCTCTATGGGCAGTCGGTGATGTAGTA	29
P1-BC-5-5′Oligo	TGTAAGATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-29-5′Oligo	CTAGGGATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-5-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCTTACA	29	P1-BC-29-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCCCTAG	29
P1-BC-6-5′Oligo	GGGACAATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-30-5′Oligo	CGAAGAATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-6-3′Oligo	CCTCTCTATGGGCAGTCGGTGATTGTCCC	29	P1-BC-30-3′Oligo	CCTCTCTATGGGCAGTCGGTGATTCTTCG	29
P1-BC-7-5′Oligo	TATGCCATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-31-5′Oligo	GTGCGTATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-7-3′Oligo	CCTCTCTATGGGCAGTCGGTGATGGCATA	29	P1-BC-31-3′Oligo	CCTCTCTATGGGCAGTCGGTGATACGCAC	29
P1-BC-8-5′Oligo	GAGGATATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-32-5′Oligo	TGGGTGATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-8-3′Oligo	CCTCTCTATGGGCAGTCGGTGATATCCTC	29	P1-BC-32-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCACCCA	29

P1-BC-9-5′Oligo	TGCGACATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-33-5′Oligo	GGATACATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-9-3′Oligo	CCTCTCTATGGGCAGTCGGTGATGTCGCA	29	P1-BC-33-3′Oligo	CCTCTCTATGGGCAGTCGGTGATGTATCC	29
P1-BC-10-5′Oligo	TAAGCTATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-34-5′Oligo	TAAAGGATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-10-3′Oligo	CCTCTCTATGGGCAGTCGGTGATAGCTTA	29	P1-BC-34-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCCTTTA	29
P1-BC-11-5′Oligo	GACACGATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-35-5′Oligo	GTAGGAATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-11-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCGTGTC	29	P1-BC-35-3′Oligo	CCTCTCTATGGGCAGTCGGTGATTCCTAC	29
P1-BC-12-5′Oligo	GGAAAAATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-36-5′Oligo	TCACATATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-12-3′Oligo	CCTCTCTATGGGCAGTCGGTGATTTTTCC	29	P1-BC-36-3′Oligo	CCTCTCTATGGGCAGTCGGTGATATGTGA	29
P1-BC-13-5′Oligo	TAAGGCATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-37-5′Oligo	TATACCATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-13-3′Oligo	CCTCTCTATGGGCAGTCGGTGATGCCTTA	29	P1-BC-37-3′Oligo	CCTCTCTATGGGCAGTCGGTGATGGTATA	29
P1-BC-14-5′Oligo	GGCAGAATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-38-5′Oligo	TCTTAGATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-14-3′Oligo	CCTCTCTATGGGCAGTCGGTGATTCTGCC	29	P1-BC-38-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCTAAGA	29
P1-BC-15-5′Oligo	TGAATGATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-39-5′Oligo	GTGAGTATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-15-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCATTCT	29	P1-BC-39-3′Oligo	CCTCTCTATGGGCAGTCGGTGATACTCAC	29
P1-BC-16-5′Oligo	GACGTTATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-40-5′Oligo	GGGTTAATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-16-3′Oligo	CCTCTCTATGGGCAGTCGGTGATAACGTC	29	P1-BC-40-3′Oligo	CCTCTCTATGGGCAGTCGGTGATTAACCC	29

(continued)

(continued)

P1-BC-17-5′Oligo	CACCGCATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-41-5′Oligo	GGATTGATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-17-3′Oligo	CCTCTCTATGGGCAGTCGGTGATGCGGTG	29	P1-BC-41-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCAATCC	29
P1-BC-18-5′Oligo	GGATGAATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-42-5′Oligo	GTACTAATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-18-3′Oligo	CCTCTCTATGGGCAGTCGGTGATTCATCC	29	P1-BC-42-3′Oligo	CCTCTCTATGGGCAGTCGGTGATTAGTAC	29
P1-BC-19-5′Oligo	TAACTGATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-43-5′Oligo	AGGGTTATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-19-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCAGTTA	29	P1-BC-43-3′Oligo	CCTCTCTATGGGCAGTCGGTGATAACCCT	29
P1-BC-20-5′Oligo	AGCTTTATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-44-5′Oligo	ATGATCATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-20-3′Oligo	CCTCTCTATGGGCAGTCGGTGATAAAGCT	29	P1-BC-44-3′Oligo	CCTCTCTATGGGCAGTCGGTGATGATCAT	29
P1-BC-21-5′Oligo	GGTTCCATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-45-5′Oligo	TGGCTCATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-21-3′Oligo	CCTCTCTATGGGCAGTCGGTGATGGAACC	29	P1-BC-45-3′Oligo	CCTCTCTATGGGCAGTCGGTGATGAGCCA	29
P1-BC-22-5′Oligo	AGATGAATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-46-5′Oligo	GTCGAAATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-22-3′Oligo	CCTCTCTATGGGCAGTCGGTGATTCATCT	29	P1-BC-46-3′Oligo	CCTCTCTATGGGCAGTCGGTGATTTCGAC	29
P1-BC-23-5′Oligo	TGCTTGATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-47-5′Oligo	GGATGGATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-23-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCAAGCA	29	P1-BC-47-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCCATCC	29
P1-BC-24-5′Oligo	CGGTATATCACCGACTGCCCATAGAGAGGTT	31	P1-BC-48-5′Oligo	TAAGTGATCACCGACTGCCCATAGAGAGGTT	31
P1-BC-24-3′Oligo	CCTCTCTATGGGCAGTCGGTGATATACCG	29	P1-BC-48-3′Oligo	CCTCTCTATGGGCAGTCGGTGATCACTTA	29

The unique barcode sequence tags are presented in color; forward and reverse oligos are presented; BC barcode.

23. Alternatively use AMPure Beads to size-select and purify adaptor-ligated DNA fragments as described in the SOLiD User Manual. In our experience, the FlashGel Recovery System (Lonza) and E-Gel SizeSelect system (Thermo Fisher Scientific) did not work as efficiently as the manual gel-based size selection.
24. Do not heat-dissolve the gel as recommended in the manufacturer's protocol to protect the DNA from denaturation and heteroduplex formation.
25. It would be possible to cut out the vector sequence with *Not1* first, and to recover insert DNA with a gel-based size selection.
26. To avoid chimerism, the DNA concentration should be low and the reaction volume high enough to increase the chances that ligation will occur between the two ends of one DNA molecule, rather than two different DNA molecules.
27. DNA polymerase I activity is highly temperature-sensitive, chilled reagents and ice bathing are crucial to create optimal tag length. Thus, a higher temperature, or increased incubation time, may lead to an extension of the size of the fragments.
28. Consistent with the SOLiD sequencing protocol, barcoded adaptors for mate-paired sequencing were not established at the time.
29. Use differently colored caps, if several fosmid pool DNA samples are processed.
30. Run as few cycles as necessary to achieve visible gel bands; over-amplification can lead to redundant library molecules and reduces the library complexity. Start with 5 cycles, if DNA concentration is about 400 ng/μl, start with up to 8 cycles, if DNA concentration is about 50 ng/μl. If necessary, pause PCR reaction and check an aliquot on a gel before running additional PCR cycles.
31. For paired-end libraries, the final concentration needs to be 50 pg/μl and for mate-paired libraries, the final concentration needs to be 96 pg/μl to correspond to 500 pM.
32. Do not magnet the P2-enriched beads before denaturing buffer has been added.
33. The use of the standard SAM or BAM format at this step, as well as the VCF format at the next step, is helpful to achieve independence from the sequencing platform for fosmid detection and haplotyping.
34. The identification of variants against the reference genome can be achieved directly from fosmid pool NGS and analysis by merging the BAM files from each pool as described here, or, alternatively, by analysis of WGS data, in the case where WGS is performed in complementation to fosmid pool NGS, *see* also Fig. 6.

35. In the latter case, if n1 and n2 are the largest and second largest allele coverage, respectively, and t1 and t2 are two corresponding minimum thresholds (with defaults 1 and 2, respectively), a homozygous FSG is called if $n1 \geq t1$ and either $n2 < t2$ or $n1 \geq 2 \times n2$. If the first condition does not hold, the FSG is left uncalled. If the second condition does not hold, a heterozygous FSG is called.

References

1. Hoehe MR (2003) Haplotypes and the systematic analysis of genetic variation in genes and genomes. Pharmacogenomics 4:547–570. doi:10.1517/phgs.4.5.547.23791
2. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ (2011) The importance of phase information for human genomics. Nat Rev Genet 12:215–223. doi:10.1038/nrg2950
3. Suk EK, McEwen GK, Duitama J, Nowick K, Schulz S, Palczewski S, Schreiber S, Holloway DT, McLaughlin S, Peckham H et al (2011) A comprehensively molecular haplotype-resolved genome of a European individual. Genome Res 21:1672–1685. doi:10.1101/gr.125047.111
4. Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, Ng SB, Alkan C, Qiu R, Eichler EE et al (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat Biotechnol 29:59–63. doi:10.1038/nbt.1740, nbt.1740 [pii]
5. Duitama J, McEwen GK, Huebsch T, Palczewski S, Schulz S, Verstrepen K, Suk EK, Hoehe MR (2012) Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of single individual haplotyping techniques. Nucleic Acids Res 40:2041–2053. doi:10.1093/Nar/Gkr1042
6. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang YT, Haas J et al (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature 487:190–195. doi:10.1038/nature11236
7. Kaper F, Swamy S, Klotzle B, Munchel S, Cottrell J, Bibikova M, Chuang HY, Kruglyak S, Ronaghi M, Eberle MA et al (2013) Whole-genome haplotyping by dilution, amplification, and sequencing. Proc Natl Acad Sci U S A 110:5552–5557. doi:10.1073/pnas.1218696110
8. Lo C, Liu R, Lee J, Robasky K, Byrne S, Lucchesi C, Aach J, Church G, Bafna V, Zhang K (2013) On the design of clone-based haplotyping. Genome Biol 14:R100. doi:10.1186/gb-2013-14-9-r100
9. Hoehe MR, Church GM, Lehrach H, Kroslak T, Palczewski S, Nowick K, Schulz S, Suk EK, Huebsch T (2014) Multiple haplotype-resolved genomes reveal population patterns of gene and protein diplotypes. Nat Commun. 2014 Nov 26;5:5569. doi: 10.1038/ncomms6569
10. Burgtorf C, Kepper P, Hoehe M, Schmitt C, Reinhardt R, Lehrach H, Sauer S (2003) Clone-based systematic haplotyping (CSH): a procedure for physical haplotyping of whole genomes. Genome Res 13:2717–2724. doi:10.1101/Gr.1442303
11. Duitama J, Huebsch T, McEwen G, Suk E-K, Hoehe MR (2010) ReFHap: a reliable and fast algorithm for single individual haplotyping. Proceedings of the first ACM international conference on bioinformatics and computational biology ACM, Niagara Falls, New York, pp. 160–169
12. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al (2015) A global reference for human genetic variation. Nature 526:68–74. doi:10.1038/nature15393
13. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. doi:10.1093/bioinformatics/btp324
14. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9:357–359. doi:10.1038/nmeth.1923
15. Duitama J, Quintero JC, Cruz DF, Quintero C, Hubmann G, Foulquie-Moreno MR, Verstrepen KJ, Thevelein JM, Tohme J (2014) An integrated framework for discovery and genotyping of genomic variants from high-throughput sequencing experiments. Nucleic Acids Res 42:e44. doi:10.1093/nar/gkt1381
16. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. doi:10.1038/ng.806
17. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter

estimation from sequencing data. Bioinformatics 27:2987–2993. doi:10.1093/bioinformatics/btr509

18. Nedelkova M, Maresca M, Fu J, Rostovskaya M, Chenna R, Thiede C, Anastassiadis K, Sarov M, Stewart AF (2011) Targeted isolation of cloned genomic regions by recombineering for haplotype phasing and isogenic targeting. Nucleic Acids Res 39:e137. doi:10.1093/nar/gkr668
19. Donahue WF, Ebling HM (2007) Fosmid libraries for genomic structural variation detection. Curr Protoc Hum Genet Chapter 5:Unit 5.20. doi:10.1002/0471142905.hg0520s54