nature biotechnology

High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing

Mark Matzas[1,5], Peer F Stähler[1,5], Nathalie Kefer[1], Nicole Siebelt[1], Valesca Boisguérin[1], Jack T Leonard[1], Andreas Keller[1], Cord F Stähler[1], Pamela Häberle[1], Baback Gharizadeh[2], Farbod Babrzadeh[2] & George M Church[3,4]

The construction of synthetic biological systems involving millions of nucleotides is limited by the lack of high-quality synthetic DNA. Consequently, the field requires advances in the accuracy and scale of chemical DNA synthesis and in the processing of longer DNA assembled from short fragments. Here we describe a highly parallel and miniaturized method, called megacloning, for obtaining high-quality DNA by using next-generation sequencing (NGS) technology as a preparative tool. We demonstrate our method by processing both chemically synthesized and microarray-derived DNA oligonucleotides with a robotic system for imaging and picking beads directly off of a high-throughput pyrosequencing platform. The method can reduce error rates by a factor of 500 compared to the starting oligonucleotide pool generated by microarray. We use DNA obtained by megacloning to assemble synthetic genes. In principle, millions of DNA fragments can be sequenced, characterized and sorted in a single megacloner run, enabling constructive biology up to the megabase scale.

Current *de novo* gene construction[1–4] rests on 1990's technology for chemical oligonucleotide synthesis, which is costly and has error rates of 1 in 300 base pairs (bp). Errors are typically avoided by manually selecting the best Sanger sequences using electrophoretic automation. Recent innovations in programmable array technology[5–8] offer the possibility to synthesize pools of thousands to millions of sequences per array with lengths comparable to conventional synthesis. The technology thus provides an extremely rich source of DNA oligonucleotides with great flexibility and superior efficiency regarding throughput and cost per bp. However, the error rate of microarray-derived oligonucleotides is typically higher compared to conventional synthesis, making error avoidance or correction necessary. Furthermore it is challenging to divide the derived oligonucleotide pools, containing vast amounts of species, into subpools—necessary, for example, to guide the assembly of synthetic genes, chromosomal regions or whole pathways in synthetic biology.

Megacloning turns NGS from a previously purely analytical method into a preparative tool, and represents a tremendous source of sequence-verified DNA where the yield from one NGS run is equivalent to that from hundreds to thousands of Sanger-sequence runs. It therefore addresses the challenge of error reduction for both conventional and microarray-derived DNA oligonucleotides. The method yields high-quality DNA libraries containing perfect parts with desired and correct sequences in adjustable ratios useful for a wide range of (bio-)technological applications.

Here we present a proof-of-concept study aimed at the retrieval of clonal DNA with known sequence from an NGS platform after sequencing (**Fig. 1**). The workflow comprises the input of DNA of short length, an NGS run to generate sequence-verified DNA clones, the identification of DNA with desired sequence on the sequencer's substrate and the retrieval of the clones of choice. The sources for the input DNA are for the most part independent of the megacloning step. For the present work, input DNA was derived from conventional oligonucleotide synthesis and from DNA microarrays. We used the NGS platform GS FLX from Roche 454 Life Sciences[9,10]. Owing to its open-top architecture, accessibility of the beads and the bead size, this platform is well suited for a pick-and-place approach using micropipettes to retrieve specific beads from the 454-Picotiterplate (PTP) and transfer them into conventional multi-well plates for further processing.

First, we established a technical setup for the controlled extraction of beads. The PTP at this stage contained a natural sample from human DNA, and extraction was done using a micropipette controlled by a microactuator device (**Supplementary Data**). To assess the fidelity of our setup, we compared the reads coming from the GS FLX platform with Sanger-derived sequences of DNA amplified from extracted beads. The alignment of Sanger sequences to the NGS reads matched 99.9%. Only two mismatches were obtained in 2,410 bp. Both were putative insertions in the GS FLX reads occurring at homopolymer stretches and therefore have a high likelihood of being platform-specific, base-calling artifacts[9] (**Supplementary Data**).

Next we collected a set of 319 beads with DNA clones from a microarray-derived pool initially containing 3,918 sequences. The beads for extraction were selected to ensure that their GS FLX reads perfectly matched sequences in our starting pool. The obtained DNA and the

[1]febit group, Heidelberg, Germany. [2]Stanford Genome Technology Center, Stanford University, Palo Alto, California, USA. [3]Harvard Medical School, Boston, Massachusetts, USA. [4]Wyss Institute for Biologically Inspired Engineering, Boston, Massachusetts, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to G.M.C. (gmc@harvard.edu).

Received 8 June; accepted 19 October; published online 28 November 2010; doi:10.1038/nbt.1710

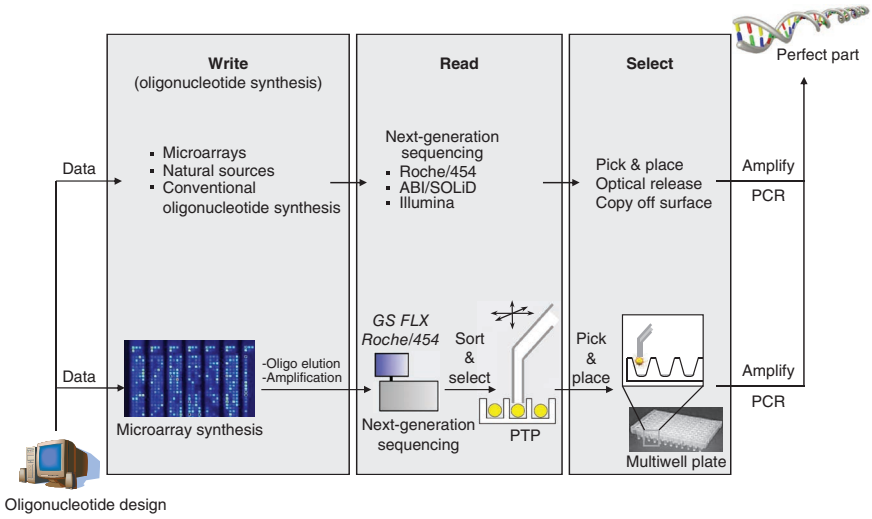

Figure 1 Coalescence of DNA reading and writing. The general approach begins with DNA from a variety of sources. Here we used oligonucleotides synthesized from microarrays as well as from conventional sources. Then, next-generation sequencing is used to read and identify oligonucleotides with desired sequences. Here we used the GS FLX platform (454/Roche). Finally, the DNA is sorted and retrieved selectively, in this case with a microactuator-controlled micropipette guided by two microscope cameras. The technologies used for retrieval depend on the sequencing platform.

untreated pool were compared after being sequenced independently on a Genome Analyzer II (Illumina GAII). We mapped 3.1% of reads from the initial (nonenriched) DNA pool without errors to the set of 319 selected sequences. In the enriched pool the fraction of reads mapping perfectly to the target sequences was 84.3%. The increase by a factor of 27.2 shows clearly a successful enrichment of selected and correct sequences (**Fig. 2a**,**b**). Also the analysis of reads on the level of single-target sequences shows that for 94% of the sequences in the selected pool, 50% or more of the reads were correct (**Fig. 2c**). Error-prone sequences contained a high number of different species likely to be caused by known sequence variations on the GAII, as reported previously[11].

To test the assembly of gene fragments based on megacloned oligonucleotides stemming from a microarray, we assembled two gene fragments, each ~220 bp in length, combining either nine or ten megacloned, bead-derived amplicons in a PCR-based gene assembly reaction[12,13]. The obtained assemblies were cloned and Sanger sequenced. Seven out of eight clones matched the target sequence perfectly. Interestingly, one clone showed insertions and deletions all located within a region 23 bp wide. Errors in assemblies originating from inaccuracies in the starting material could be expected to be distributed evenly over the entire construct. As this sequence was otherwise free of errors, these defects were likely caused by misassembly rather than errors in the building blocks used (**Supplementary Data**).

To further evaluate the capabilities of the megacloning approach to generate biologically functional genes, we applied the method to DNA fragments 274–394 bp in length and extracted 32 beads from the PTP carrying putatively correct sequences. These DNA fragments were the product of gene assembly reactions[12] using overlapping 40-mer oligonucleotides synthesized using conventional phosphoramidite chemistry and could be assembled into a model gene encoding β-D-glucuronidase (*uidA*)[14] (2,080 bp).

Three Sanger sequences obtained from the bead DNA were totally unrelated to the expected sequence and were probably caused by wrong bead extraction or contamination. The remaining 29 sequences covered 7,195 bp and matched without errors to the expected target sequences (**Supplementary Data**).

We then assembled the model gene out of nine DNA fragments from the set of 29 matching beads. The full-length gene construct was again checked by Sanger sequencing for absence of errors, and the biological functionality of the gene was tested in an enzymatic assay based on the conversion of X-Glc (5-bromo-4-chloro-3-indolyl-β-glucoside) substrate into blue dye[15] (**Supplementary Data**). Besides the proof of feasibility of generating biological functional genes, this experiment further mimics other applications of our technology, such as the use of sheared natural DNA and its subsequent sorting and reordering.

The absence of errors in 7,195 bp of DNA obtained from 29 extracted beads raised the question of achievable error rates from the megacloner process. Therefore we explored the potential of megacloning using a statistical model. This model considers two main sources of error—namely, wrong sequencing calls and polymerase errors during DNA amplification[16]. The calculations estimated the chance of finding one error in our extracted sequence space of ~7,200 bp to be 29%, which is in line with our experimental findings. The theoretical error rate of bead amplicons after megacloning using the setup employed in this study was estimated to be 1 error in 21 kbp (**Supplementary Data**). Compared with the error rate in the starting material of 1 error in 40 bp (determined from GAII data of the initial microarray pool), this equals a 500-fold error reduction.

We further calculated the expected amount of reads from NGS that match the target sequences of a given pool without errors. These numbers are crucial to estimate the complexity of pools that can be processed in one megacloner run. The resulting efficiency and cost structure are influenced mainly by three parameters: the error rate of the starting pool, the sequencing accuracy and the length of the variable sequence (**Supplementary Data**). With an error rate of 1 error in 40 bp and an average sequencing accuracy of 99.9% in the GS FLX, we expect a five- to tenfold cost reduction in producing DNA fragments (compared to conventional oligonucleotide synthesis) that can be achieved now with the prototype device (**Supplementary Data**). Because these fragments are largely free of errors, further savings can be expected in gene synthesis because the cost of subsequent sequencing for final quality control will be lower.

In this work we demonstrated the targeted retrieval of bead-bound DNA from a high-throughput sequencer without major modifications to the sequencing process. Previous methods for error correction in DNA pools[7,17–21] do not adequately handle collections of closely related oligonucleotide sequences that occur during assembly of repetitive sequences or multi-gene family libraries. They also do not enable hierarchical assembly strategies, which are made possible by the ordered selection and physical separation of clonal DNA described here.

The megacloner process has been proven to be useful for retrieval and sorting of correct and functional sequences and to increase the portion of error-free sequences in a sample substantially. This technology allows the processing of DNA from microarrays but also from a variety of other sources, such as conventional oligonucleotide synthesis or natural DNA fragments.

Megacloning could be optimized beyond the estimates in this work of one error in 21 kbp from input DNA having an error rate of 1 in 40 bp. Although such raw material can be obtained by state-of-the-art microarray technologies, the quality of input DNA could be increased further by addressing the amplification step of bead-bound DNA—for example, with higher fidelity polymerases, as the predicted contribution of the polymerase to the error rate is 4.7-fold higher than the expected error rate of the megacloner itself (**Supplementary Data**). Another accessible parameter for optimizing the overall process in terms of error rates is improvement in the quality of the DNA starting material. Also, optimization of sequencing accuracy could be a way to improve the ability to select correct parts after NGS. This is, however, the subject of ongoing optimization in the scope of NGS development, including ligase-based methods with improved accuracy[22].

The pool used in our conceptual study contained ~4,000 sequences. According to our results and extrapolations, this can be increased to ~30,000 sequences per pool with the described setup. As the bead extraction is generally independent of the pool complexity, it is mainly limited by the NGS platform and the quality of the starting material (**Supplementary Data**). More advanced microarray formats are able to deliver libraries with even higher complexity and of sufficient quality to fit into a gene assembly process[23]. Therefore, with an appropriate degree of automation that reaches an extraction frequency of two or three beads per minute, which is achievable with state-of-the-art robotics, the work-up of one PTP becomes possible within days, resulting in > 10^6 bp per plate. Hence, the downstream process (amplification, cleanup, assembly) will represent the next bottleneck.

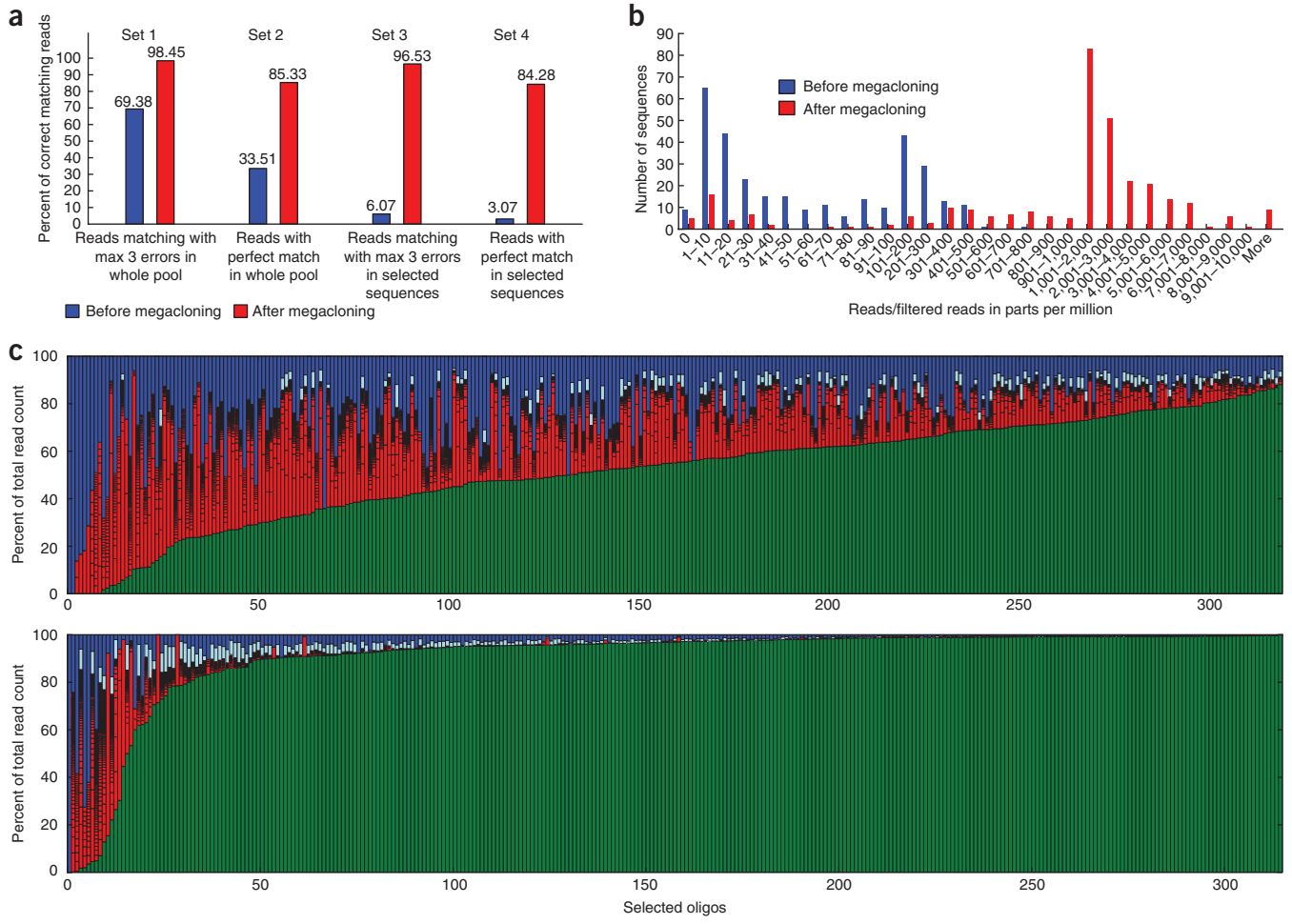

Figure 2 NGS-based comparison of untreated and megacloned oligonucleotide pools from microarray. (**a**) Comparison of the initial microarray oligonucleotide pool (blue) and the pool enriched with the megacloner technology (red) based on the results of the Illumina GAII runs. The bars in set 1 represent the fraction of reads that could be mapped allowing up to three errors. Bars in set 2 show the fractions of perfectly matching reads to the sequence set of the initial pool (3,918 sequences). The difference between the blue and the red bar in set 2 represents the enrichment of correct sequences by megacloning. The bars in set 3 and set 4 show the fractions of reads mapping to sequences from the selected pool of 319 sequences. The difference between blue and red bars in set 3 shows the enrichment of a selected 319 sequences before megacloning compared with after. Blue and red bars in set 4 represent the enrichment of sequences that are in the set of 319 selected sequences and that are correct. (**b**) Histogram of read counts in the Illumina GAII data of the initial pool (blue) and the enriched megacloned sample (red). Only reads mapping without errors to one of the 319 selected target sequences have been taken into account. To compare the two NGS runs on the basis of read counts, we converted the numbers into parts-per-million (p.p.m.) from the total number of filtered reads. (**c**) Composition of reads from the Illumina GAII data including 319 selected sequences in the initial pool (top) and the enriched pool (bottom). The oligonucleotides are sorted by the fraction of correct reads. Green, correct reads; red, error-prone reads (compartments in the red bars represent single sequences with a read count of 0.1% or more of total reads for the particular sequence); light blue, sum of nonunique error-prone reads where each sequence represents less than 0.1% of total reads for the particular sequence; blue, unique reads. In the Illumina GAII data set from the enriched sample, just 315 out of 319 selected sequences could be detected.

Our next focus in the present context is improvement and automation of physical bead extraction. The workflow used in this study still involved a considerable number of manual steps and some human intervention, which was identified as the most important source of error in terms of extraction of unwanted beads. Therefore, the success rate of ~90% (29 beads out of 32) has to be increased for the bead localization and retrieval process.

The method described here holds the potential to decrease production cost for synthetic DNA by one or more orders of magnitude. This source of high-quality DNA could aid the field of synthetic biology, as well as the production of libraries for antibodies or enzyme variants. In addition to synthetic sources, the sorting of natural DNA could enable the quick reconstruction or combination of DNA fragments to assemble genes, chromosomes or genomes, while simultaneously including synthetic parts of DNA.

The principle that we applied here using the GS FLX technology should also be generally applicable to other available NGS platforms such as Illumina's GAII, SOLiD, the Polonator or others. In the present context, the advantage of the GS FLX platform is the robot-accessible platform architecture and the comparably large size of the beads. Owing to different architectures of the other platforms, such as partially closed systems and substantially smaller DNA carriers, harvesting DNA from those will require a different mechanism, such as optical approaches including photosensitive and cleavable linker-molecules. The advantage of these platforms is a considerably higher number of DNA clones, which potentially could increase the capacity and throughput of the technology up to the gigabase level.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank B.A. Roe, F.Z. Najar and D.D. White for sequencing support, J. Jäger for technical consulting, and D. Summerer, T. Brefort, S. Kosuri and D. Levner for discussions and comments.

AUTHOR CONTRIBUTIONS

M.M., P.F.S. and G.M.C. conceptualized the megacloning method and wrote the manuscript; M.M. designed and lead the study, wrote all algorithms for sequence design, data analysis, image conversion, image processing and microactuator control; M.M., N.K., N.S. acquired the used technology, set up the microactuator device and optical systems; N.S. designed the *uidA* genetic model; M.M., N.K., N.S., V.B. and P.H. designed and optimized molecular biological methods; C.F.S. and J.T.L. contributed to bead picking and engineering concepts; A.K. set up the statistical models and calculations; J.T.L. contributed to the design of molecular biological steps and the acquisition of sequencing samples; B.G. and F.B. evaluated and implemented necessary changes into the sample preparation and the sequencing process on the 454/Roche platform.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Endy, D. Foundations for engineering biology. *Nature* **438**, 449–453 (2005).
2. Menzella, H.G. *et al.* Combinatorial polyketide biosynthesis by *de novo* design and rearrangement of modular polyketide synthase genes. *Nat. Biotechnol.* **23**, 1171–1176 (2005).
3. Gibson, D.G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).
4. Carr, P.A. & Church, G.M. Genome engineering. *Nat. Biotechnol.* **27**, 1151–1162 (2009).
5. Gao, X. *et al.* A flexible light-directed DNA chip synthesis gated by deprotection using solution photogenerated acids. *Nucleic Acids Res.* **29**, 4744–4750 (2001).
6. Singh-Gasson, S. *et al.* Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.* **17**, 974–978 (1999).
7. Tian, J. *et al.* Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* **432**, 1050–1054 (2004).
8. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
9. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
10. Wicker, T. *et al.* 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7**, 275 (2006).
11. Willenbrock, H. *et al.* Quantitative miRNA expression analysis: comparing microarrays with next-generation sequencing. *RNA* **15**, 2028–2034 (2009).
12. Stemmer, W.P., Crameri, A., Ha, K.D., Brennan, T.M. & Heyneker, H.L. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **164**, 49–53 (1995).
13. Richmond, K.E. *et al.* Amplification and assembly of chip-eluted DNA (AACED): a method for high-throughput gene synthesis. *Nucleic Acids Res.* **32**, 5011–5018 (2004).
14. Jefferson, R.A., Burgess, S.M. & Hirsh, D. beta-Glucuronidase from *Escherichia coli* as a gene-fusion marker. *Proc. Natl. Acad. Sci. USA* **83**, 8447–8451 (1986).
15. Couteaudier, Y., Daboussi, M.J., Eparvier, A., Langin, T. & Orcival, J. The GUS gene fusion system (*Escherichia coli* beta-D-glucuronidase gene), a useful tool in studies of root colonization by Fusarium oxysporum. *Appl. Environ. Microbiol.* **59**, 1767–1773 (1993).
16. Cline, J., Braman, J.C. & Hogrefe, H.H. PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.* **24**, 3546–3551 (1996).
17. Carr, P.A. *et al.* Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Res.* **32**, e162 (2004).
18. Smith, J. & Modrich, P. Removal of polymerase-produced mutant sequences from PCR products. *Proc. Natl. Acad. Sci. USA* **94**, 6847–6850 (1997).
19. Bang, D. & Church, G.M. Gene synthesis by circular assembly amplification. *Nat. Methods* **5**, 37–39 (2008).
20. Fuhrmann, M., Oertel, W., Berthold, P. & Hegemann, P. Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage. *Nucleic Acids Res.* **33**, e58 (2005).
21. Binkowski, B.F., Richmond, K.E., Kaysen, J., Sussman, M.R. & Belshaw, P.J. Correcting errors in synthetic DNA through consensus shuffling. *Nucleic Acids Res.* **33**, e55 (2005).
22. McKernan, K.J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).
23. Kosuri, S. *et al.* Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat. Biotechnol.* advance online publication, doi:10.1038/nbt.1716 (28 November 2010).

ONLINE METHODS

Oligo synthesis, sequence design, adaptors. Oligonucleotides used for this work were synthesized on programmable microarray synthesizers using light-directed synthesis methods[5]. Conventional oligonucleotides used for gene assembly were obtained from Sigma Aldrich. Harvesting of oligonucleotides from microarray surfaces was performed by chemical cleavage of succinate-ester bonds using ammonia hydrochloride solution.

Amplification of microarray-derived oligonucleotide pools by emulsion PCR. Microarray-derived oligonucleotide pools were amplified before NGS using emulsion PCR[24]. Therefore universal terminal sequences were attached during synthesis and served as primer regions. Amplification primers contained adaptors for sequencing on the Illumina GAII platform and/or the 454 GS FLX (**Supplementary Data**).

Sequencing on the 454 GS FLX. The sample preparation for the PCR-amplified oligonucleotides was done according to the manufacturer's protocols (Roche/454). To keep the DNA intact after sequencing, we exchanged the bleaching cleaning buffer with TE buffer before the sequencing run to avoid degradation of DNA during the final cleaning steps of the Roche sequencer.

Data analysis of 454 data and image conversion. NGS reads obtained from the GS FLX sequencer were aligned to the target sequences in the oligonucleotide pool to find the best matching sequence for every read and to perform further analysis, such as error rate estimation. Perfect matching sequences were selected and localized in the sequencer image by using the coordinates attached to every read sequence. For sequence data analysis, we used various Python scripts using the BioPython package. The images from the GS FLX sequencer were converted into the TIFF standard format using the Python Imaging Library.

Bead localization and extraction. After aligning the GS FLX reads to the set of target sequences, we selected reads that perfectly matched one of the desired oligonucleotide sequences in the pool. For localization of beads we located the corresponding chemiluminescent signals in the converted raw image from the GS FLX platform using the *x*- and *y*-coordinates that were included in the NGS raw data. To locate beads in the PTP, we identified reference points in the raw image and their corresponding positions in the PTP using suitable patterns of light signals. Based on these reference points the bead positions on the PTP were calculated using an algorithm for scaling and rotation. The extraction was performed with a micropipette with an outer diameter of 28 µm. For pipette handling we used a three-axis microactuator (**Supplementary Data**). Before extraction of beads the PTP was stored under a water layer to prevent desiccation and shrinking of beads. After picking, the beads were transferred immediately into a PCR vial and stored under water until further processing.

Amplification of DNA from beads. Amplification of bead-bound DNA was performed with the primers 454-A and 454-B, targeting the Roche/454 adaptors, or 'slx-fw-long' and 'slx-rev-long' for Illumina adaptors. For amplification of fragments with 40-mer variable regions, primers were 5′-biotinylated to facilitate subsequent removal of primer regions on a streptavidin matrix. PCR conditions: 20 mM Tris-HCl (pH 8.8), 10 mM ammonium-sulfate, 10 mM potassium chloride, 2 mM magnesium-sulfate, 0.1% Triton X-100, 200 µM each dNTP, 2% (vol/vol) DMSO, 1 µM each primer, 50 U/ml native pfu polymerase (Fermentas). Cycling: initial denaturation 96 °C (2 min); then 30 cycles of 96 °C (30 s), 63 °C (30 s), 72 °C (30 s) and final elongation 72 °C (3 min). After amplification, all PCR products were analyzed on PAGE (**Supplementary Data**) to check specificity and yield. For generation of the subpool containing 319 sequences, we estimated the concentration on the basis of the gel analysis and mixed the amplicons in equimolar concentrations.

Illumina sequencing and data analysis. As the sample contained suitable adaptors all steps regarding adaptor ligation have been omitted. All other steps were done according to the protocols from Illumina.

The NGS raw data obtained from Illumina GAII were processed by the following steps.

1. Truncation of reads to the length of the variable regions (40 bp).
2. Filtering out reads containing ambiguities (filtered reads).
3. Group reads with similar sequences (bins).

Subsequently for each read we identified the best matching target sequence from the oligonucleotide pool by mapping all reads to a pseudo-genome using rapid alignment of small RNA reads (razerS) (http://www.seqan.de/projects). The pseudo-genome was generated by concatenation of the variable parts of pool sequences separated by 40-mer poly-T stretches. The corresponding target sequence could then be determined by the matching position in the pseudo-genome. Alignments from the razerS output were used to determine insertions, deletions and substitutions. To compare the two GAII runs based on the number of correct reads, we converted the read counts into parts-per-million units (p.p.m.), taking the number of filtered reads before the matching procedure (after step 2) as a basis.

Assembly of gene fragments from conventional oligonucleotides. Gene fragments > 200 bp were assembled from conventionally synthesized 40-mer oligonucleotides having a constant overlap region of 20 nucleotides to the adjacent oligomer. Primer regions for 454 sequencing and restriction sites for primer removal were included during assembly. The assembly reaction contained 5 nM of each construction oligonucleotide and 200 nM of terminal primers. PCR conditions: 1× KOD polymerase buffer (Novagen), 1.25 mM $MgSO_4$, 40 µM each dNTP, 5 U/ml KOD Hot Start Polymerase (Novagen). Cycling for gene assembly: initial denaturation 96 °C (4 min); then 30 cycles of 96 °C (10 s), 55–40 °C touchdown (30 s), 72 °C (10 s). For subsequent amplification: 96 °C (10 s), 55 °C (30 s), 72° (30 s), final elongation 72 °C (3 min).

Assembly of genes from >200 bp fragments. Gene assembly up to 2 kbp were performed according to the protocol used for assembly of > 200 bp from oligonucleotides.

Primer removal and cleanup of bead amplicons before gene assembly. For removal of primer regions amplicons were incubated with LguI restriction endonuclease in 1× Tango buffer (Fermentas) for 3 h at 37 °C. For > 200 bp fragments, small restriction fragments containing primer regions were removed by PCR purification columns (GenElute PCR Clean-Up, Sigma Aldrich). For cleanup of microarray-derived fragments, we used 40-mer variable region biotinylated primers during bead DNA amplification and removed restriction products containing biotin residues using streptavidin matrix. The 40-mer fragments were ethanol precipitated and dissolved in water before further processing.

Assembly of genes from 40-mer double-stranded DNA fragments. For the assembly of genes from 40-mer dsDNA we used a two-stage assembly protocol including a primerless PCR followed by a PCR for amplification of the resulting products described previously[13].

24. Williams, R. *et al.* Amplification of complex gene libraries by emulsion PCR. *Nat. Methods* **3**, 545–550 (2006).