ACCURATE ASSESSMENT OF Tn5-BASED DNA FRAGMENTATION EFFICIENCY AND EVALUATION OF TAGMENTATION SITE INDEXING APPROACH FOR CONTIGUITY-PRESERVING SEQUENCING

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)
submitted to the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by
VERA N. RYKALINA
from Moscow, Russia

2017

1st Reviewer: Prof. Dr. Hans Lehrach

2nd Reviewer: Prof. Dr. Rupert Mutzel

Date of the disputation: 16.06.2017

RELATED PUBLICATION

Rykalina, V. N., Shadrin, A. A., Lehrach, H. and Borodina T. A. (2017) qPCR based characterization of DNA fragmentation efficiency of Tn5 transposomes. *Biol Method Protoc* 2 (1): bpx001.

OTHER PUBLICATIONS

Andreeva, T. V., Tyazhelova, T. V., **Rykalina, V. N.**, Gusev, F. E., Goltsov, A. Y., Zolotareva, O. I., Aliseichik, M. P., Borodina, T. A., Grigorenko, A. P., Reshetov, D. A., Ginter, E. K., Amelina, S. S., Zinchenko, R. A. and Rogaev, E. I. (2016) Whole exome sequencing links dental tumor to an autosomal-dominant mutation in ANO5 gene associated with gnathodiaphyseal dysplasia and muscle dystrophies. *Sci Rep* 6, 2644.

Rykalina, V. N., Shadrin, A. A., Amstislavskiy, V. S., Rogaev, E. I., Lehrach, H. and Borodina, T. A. (2014) Exome sequencing from nanogram amounts of starting DNA: comparing three approaches. *PLoS One*, 9 (7): e101154.

Borodina, T. A., **Rykalina, V. N.** and Lehrach, H. Method for Generating of Oligonucleotide Arrays Using In Situ Block Synthesis Approach. European patent PCT/EP2014/076888

Contents

ACKNOWLEDGMENTS

I cordially thank Prof. Dr. Hans Lehrach for support, inspiring discussions and generously shared knowledge. His enthusiasm, constant interest in the project and immense methodological experience made this work possible.

I thank Prof. Dr. Rupert Mutzel for reading my thesis and for his help in inviting the dissertation committee members.

I would like to express my profound gratitude and especially thank Dr. Tatiana Borodina, my project leader and a close friend. I thank her for the professional support throughout the project and valuable comments on my thesis. I learned a lot from her and I will always be immensely grateful for her advice and support.

I thank Dr. Alexey Shadrin, my friend and a colleague, for invaluable help with bioinformatics analysis and constructive suggestions on experimental design.

I am very grateful for Dr. Alisa Fuchs's advice and assistance in preparing in-house Tn5.

I express my gratitude to Dr. Marie-Laure Yaspo and her group for sequencing the PITS libraries on MiSeq platform and for the access to the group equipment.

I kindly thank the scullery team for their professional work and help.

I thank my friends Assit. Prof. Vincent Emmeline Sichula, Katerina and Steve Brumpton for proofreading this thesis.

I thank Jeannine Wilde for translating the Summary section into German.

I thank my friends and colleagues for making my PhD life easier. Encouraging and supporting me, giving an advice – they inspired me to always move further. I would also like to thank Dr. Irene Pakuscher for providing me with the comfortable and homely accommodation in Berlin.

Finally, I wish to pay tribute to my beloved family for their support and understanding.

ABBREVIATIONS

amol	attomole
BAC	bacterial artificial chromosome
bp	base pair
FEA	fragmentation efficiency assay
HLA	human leukocyte antigen
HMW	high molecular weight
kb	kilobase
LB	lysogeny broth
MALBAC	multiple annealing and looping based amplification cycles
MDA	multiple strand displacement amplification
NA	nucleic acid
NGS	next generation sequencing
nt	nucleotide
PAGE	polyacrylamide gel electrophoresis
PITS	paired indexing at tagmentation sites
PTULI	post-tagmentation ultra low input
RT	room temperature
SDS	sodium dodecylsulfate
SOC	super optimal broth with Catabolite repression
ssDNA	single-stranded DNA
STRs	short tandem repeat
T4 PNK	T4 Polynucleotide Kinase
TGS	third generation sequencing
Tsome	Tn5 transposome
Tson	Tn5 transposon
VNTRs	variable number tandem repeat

SUMMARY

A novel contiguity-preserving sequencing approach was developed and tested – paired indexing at tagmentation sites (PITS). The idea of the method is to use paired indices to individually label pairs of DNA molecule ends originating at tagmentation sites. Indices from the same pair on the ends of two sequencing library fragments would mean that those fragments were next to each other in the original molecule. Thus after getting separated in the way it occurs in a standard sequencing library preparation protocol library, fragments after sequencing would be assembled again and the contiguity of the sequence would be restored. Convenient system for testing of the PITS method was chosen and proof-of-principle experiment was performed. Though few, scaffolds containing up to 4 subsequent library molecules were assembled. Some of the constraints of the PITS approach were revealed and optimization possibilities for future work were determined. A patent application describing the PITS protocol is in preparation.

During the course of this PhD, an efficient method for preparation of sequencing library from amol amounts of tagmentation products has been established – post tagmentation ultra low input (PTULI) protocol. The method aims at preserving possibly all tagmentation fragments throughout sequencing library preparation process. The developed protocol provides detailed instructions on the controls setup and guidelines for subsequent non-residual loading of the PTULI library on a sequencer. The PTULI library preparation strategy is suitable for PITS, and is also of value for other minute input material sequencing applications.

To make contiguity-preserving sequencing work possible, in-house Tn5 transposase was prepared. Applications of this enzyme within the settings other than those in commercially available kits are hardly known and further development of the PITS approach to a great extent depends on the potential of this enzyme. That is why in parallel to the PITS method itself, Tn5 properties were studied. An electrophoresis free assay for characterizing Tn5 transposomes fragmentation efficiency was developed and published [Rykalina et al., 2017]. This assay is a convenient monitoring system for the setup of tagmentation protocols and for optimization experiments.

Zusammenfassung

Es wurde ein neuartiger Kontiguität-bewahrender Sequenzierungsansatz entwickelt und getestet: gepaarte Indizierung an Tagmentationsstellen (*Paired Indexing at Tagmentation Sites - PITS*). Die Idee der Methode besteht darin, die Paare der DNA-Molekül-Enden, die ursprünglich von Tagmentationsstellen stammen, individuell mit gepaarten Indizes zu markieren. Gleiche Indizes an den Enden der zwei sequenzierten Bibliotheksfragmente würde bedeuten, dass diese Fragmente im ursprünglichen Molekül nebeneinander angeordnet waren. Dank dieser Methode werden die Fragmente, die während der Herstellung von Sequenzierungsbibliotheken getrennt wurden, wieder zusammengesetzt und damit die Kontiguität der Sequenz der ursprünglichen DNA-Moleküle wiederhergestellt. Es wurde ein praktisches System für die Prüfung der PITS-Methode gewählt und ein *proof-of-principle* Experiment durchgeführt. DNA Ketten mit bis zu vier nachfolgenden Bibliotheksmolekülen wurden zusammengebaut. Einige der Einschränkungen des PITS-Ansatzes wurden aufgedeckt und Optimierungsmöglichkeiten für zukünftige Arbeiten ermittelt. Eine Patentanmeldung, die das PITS-Protokoll beschreibt, ist in Vorbereitung.

Im Rahmen der Arbeit wurde ein effizientes Verfahren zur Herstellung von Sequenzierungsbibliotheken aus amol-Mengen von Tagmentationsprodukten etabliert, das *Post-Tagmentation Ultra Low Input* (PTULI) Protokoll. Die Methode zielt darauf ab, möglichst alle Tagmentierungsfragmente während des gesamten Prozesses zur Vorbereitung der Sequenzierungsbibliotheken zu bewahren. Das entwickelte Protokoll enthält detaillierte Anweisungen zum Steuerungs-Setup und Richtlinien für die anschließende komplette Beladung der PTULI-Bibliothek auf einem Sequenzer. Die PTULI-Bibliotheksvorbereitungsstrategie eignet sich nicht nur für PITS, sondern auch für andere Sequenzierungsanwendungen mit geringen Input.

Um dieses Projekt zu ermöglichen, wurde eine eigene Tn5-Transposase hergestellt. Anwendungen dieses Enzyms in anderen *settings* als in handelsüblichen Kits sind kaum bekannt und die Weiterentwicklung des PITS-Ansatzes hängt weitgehend von dem Potenzial dieses Enzyms ab. Deshalb wurden parallel zur PITS-Methode selbst Tn5-Eigenschaften untersucht. Ein elektrophoresefreier Assay zur Charakterisierung von Tn5-Transposomen Fragmentierungseffizienz wurde entwickelt und veröffentlicht [Rykalina et al., 2017]. Dieser Assay bietet ein bequemes Monitoring-System für den Aufbau von Tagmentationsprotokollen und für Optimierungsexperimente.

INTRODUCTION

Sequencing gaps: loss of contiguity information

The last ten years have seen a rapid progress in speed, throughput and cost-efficiency of DNA sequencing [Pareek et al., 2011; Morey et al., 2013]. Next generation sequencing (NGS) technologies have been firmly established in academic and clinical research, revolutionizing our possibilities to sequence any nucleic acids (NAs) in living organisms, discover individual variations, link genetic information to phenotypic traits [van Dijk et al., 2014; Hurd et al., 2009].

Interestingly, the same features of NGS which turned it into a routine technology - massively parallel processing of billions of DNA fragments in a single reaction and short-read sequencing – are the reasons for the incompleteness of the obtained sequencing information. Currently, the established strategy for the preparation of NGS libraries is to start with DNA amounts corresponding to many cells and to fragment original long NA molecules into <1000bp fragments. The fragments then all together undergo enzymatic steps receiving platform specific flanking parts. From the pool of the fragments only a small random portion (in the range of <1/50000 of all library molecules) is actually sequenced. So, during sequencing procedure the contiguity of original long NA molecules is lost. Fragment relations may be to a great extent restored during subsequent sequencing analysis through building contigs of overlapping reads and/or positioning reads on a reference sequence. However, with all the computational efforts, both *de novo* sequencing and resequencing applications currently lead to an incomplete genome assembly and inaccurate scoring of individual variation [Sohn and Nam, 2016; Chaisson et al., 2015].

The reason for assembly gaps is that not all nucleotide sequences along a complete length of genomes are unique. Genomes harbour various identical, or only slightly different, sequences located next to each other or originating from different parts of the genomes; the reads falling inside of those sequences cannot be univocally assigned to the particular copy of the sequence. Repetitive sequences may be comparatively long like segmental duplications which constitute about 3% of the human genome [Alkan et al., 2011; Chaisson et al., 2015]. These are > 10kb sequencing stretches consisting of identical or near-identical sequence blocks, characterized by extreme genetic diversity and which are prone to recurrent mutations and structural rearrangements [Genomes Project Consortium, 2013; Sudmant et al., 2010]. Shorter repetitive blocks, such as STRs, VNTRs, centromeric satellite repeats, also add to ambiguities in genome reconstruction. It is impossible to determine a correct overlap in sequencing reads

having a different number of repeats. Due to tandem repeats centromere, acrocentric and secondary constrictions of chromosomes are not included in the standard reference genome.

Analyzing diploid (or polyploid) organisms, where the whole genome sequence is duplicated, adds to the necessity to deal with allelic variants. Parental genome copies (haplotypes) may vary by hundreds of kilobases of paralogous sequences which are present in different copy number and in different orientation [Boettger et al., 2012; Sudmant et al., 2010; Antonacci et al., 2014; Pyo et al., 2013]. Allelic variants might be difficult to discover, especially in extremely divergent regions, because certain haplotypes might be preferred during an assembly process [Zody et al., 2008; Raymond et al., 2005]; allelic variants may be also confused with similar but not identical repetitive sequences and *vice versa*.

Besides sequencing gaps, which are known of being unresolved, there are also the so-called muted gaps in the reference genome. These are the regions that are considered to be completed in an assembly but appear to be different in the vast majority of individuals [Eichler 2001]. Most are the result of errors in the assembly process of repeats and duplications (also allelic variants) which are collapsed or truncated. Some errors originate from experimental procedure: for example in clone-based sequencing, the sequences that are toxic to bacteria are deleted during the cloning process, which leads to deletions in the assembly. More than 2600 muted gaps are estimated to be in the human genome assembly [Chaisson et al., 2015].

Limitations with respect to reconstruction of high homology genomic regions, regions with repetitive sequences, and regions with multiple structural rearrangements lead to inconsistency of functional analyses. Discovery of regions associated with certain traits is complicated or made impossible. Among examples are monogenic diseases for which the causative variants could have been found years ago, if the genome assembly was full. Only targeted studies and accurate local *de novo* assembly have shown that medullar cystic kidney disease type 1 (MCKD1) is associated with insertion within a repetitive sequence [Kirby et al., 2013]. Similarly, for amyotrophic lateral sclerosis (ALS) it has been found that an expanded hexanucleotide repeat is responsible for 40% of cases of the disease [Renton et al., 2011; De Jesus-Hernandez et al., 2011]. Contraction of VNTR repeat, in conjunction with a point mutation within duplicated DNA, has been identified as the reason for facioscapulohumeral muscular dystrophy [Lemmers et al., 2012]. In some cases, genes and functional genomic elements have not been annotated so far. The reason for thyrotoxic hypolalaemic periodic catalysis was discovered only when the gene, which was thought to be associated with the disease, was shown to have a duplicated copy, and that particular copy

harbored mutation leading to disease phenotype. [Ryan et al., 2010]. These examples demonstrate high expectations of functional studies of the so far missing genomic information. According to the recent estimates, up to 150Mb of euchromatic genomic regions are inaccessible to standard variation analyses [Chaisson et al., 2016].

Certain haplotypes are already known to be associated with clinical traits, for example haplotypes of the apolipoprotein gene cluster may influence plasma triglyceride concentrations and the risk towards atherosclerosis [Groenendijk et al., 2001]; haplotypes of β-2 adrenergic receptor correlate with responses to drug treatment of asthma [Drysdale et al., 2000]. Besides clinical diagnostics, other fields would also benefit from resolved haplotypes [Clark, 2004]. Several articles show that dense haplotype data might be highly helpful for studying the size and structure of human populations [Lawson et al., 2012; Schiffels and Durbin, 2014]. Unique nucleotide content associated with each of two homologous copies of a chromosome is used to investigate demographic history, patterns of human migration and population bottlenecks [Sabeti et al., 2007; Vernot and Akay, 2014; Sankararaman et al., 2014]. The research published in 2010 [Green et al., 2010] demonstrated that accurate comparison of homologous chromosomes can facilitate evolutionary studies of genomes across species. Within human population greater magnitude of differentiation can be achieved by using haplotype information [Nievergelt et al., 2007]. Reconstruction of haplotypes requires not only a correct determination of allelic variants, but also assigning them to a particular parental sequence. For short read technologies, phase information about allelic variants in two nonadjacent regions is usually not recoverable [Yang et al., 2011].

Above examples vividly confirm that contiguity-preserving sequencing, providing complete genotype and haplotype resolved data, is highly demanded by applied science. Another point - quite illustrative to the contemporary ideological progress - this data is required not only for the reference genome assembly, but also resequencing of individual genomes. Growing number of biological studies supports the idea that complete interpretation of diploid genomes is not possible without unique content of two homologues chromosome sets [Tewhey et al., 2011; Glusman et al., 2014; Bansal et al., 2011]. Recent large scale genomic projects (e.g, 1000 Genomes Project Consortium, 2010 and 2012) have shown that single-nucleotide variations and structural differences between individuals are much more abundant than it was thought before [Pendelton et al., 2015]. Besides, as was mentioned above, some genomic parts are so variable, that even a high-quality reference sequence is insufficient. Thus, ideally, most informative solution would be to perform *de novo* sequencing for each individual genome.

Contiguity preserving sequencing feasible for the high throughput requires a very robust and cost-efficient technological solution, - a real challenge for methodology and bioinformatics specialists.

Solutions for contiguity preserving sequencing

At present, there are two main directions for the development of whole-genome sequencing with an affordable level of contiguity. The first direction is purely reliant on the advancements of novel long-read sequencing technologies. The second direction is driven by search of new experimental solutions for the conventional short-read sequencing. Current long-read technologies already proved to efficiently close a considerable number of gaps [Chaisson et al., 2015]. However, they are still far from being as robust as current NGS working horse – Illumina sequencing platform, and are characterized by a considerably higher price per nucleotide and a lower throughput. That is why it is also considered that though long read technologies are continuing to be developed, it may remain that the best technologies, in terms of cost per base, are read length limited. Therefore, it might prove more feasible to obtain contiguity information using accompanying techniques – as it was for Sanger technologies years ago [Schwartz et al., 2012]. For Sanger sequencing, which has been the gold standard for DNA sequencing for decades, original long DNA molecules of genomic DNA or cDNA are also fragmented to prepare sequencing templates. The problem of contiguity restoration is solved by additional efforts, e.g. hierarchical sequencing of large cloned genomic entities (fosmids, BACs); obtaining mate-pair reads from plasmid sequencing templates.

Below we describe contiguity preserving sequencing strategies – both long-read procedures and solutions compatible with current well established NGS protocols for library preparation and available read length. We also describe the methodological background of the project and its place within the whole ensemble of technological innovations in the field.

Extending the read length: single molecule sequencing

There are great expectations for the improved *de novo* sequencing performed on third-generation sequencing (TGS) platforms, which can use significantly longer templates and generate longer reads (up to 50kb - Pacific Biosciences). Third-generation sequencing and mapping technologies are a new start for another level of data, with a goal to fill many, if not all, open gaps in human genomics, which still cannot be reached by currently available methods in terms of price and quality.

The main characteristic which defines third-generation sequencing technologies and distinguishes them from the previous generation is an ability to sequence a single NA molecule in real time, while completely excluding clonal amplification step prior to sequencing [Heather and Chain, 2016]. At the time of writing, there are two commercially available TGS systems on the market, which follow this logic: Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) sequencing (www.pacificbiosciences.com) and the Oxford Nanopore Technologies sequencing platform [Haque et al., 2013]. Both technologies are able to produce the mean read lengths up to 10000-15000bp at a very high throughput. The details of chemistries used in these sequencing platforms were well reviewed, for example, by Reuter et al. [Reuter et al., 2015].

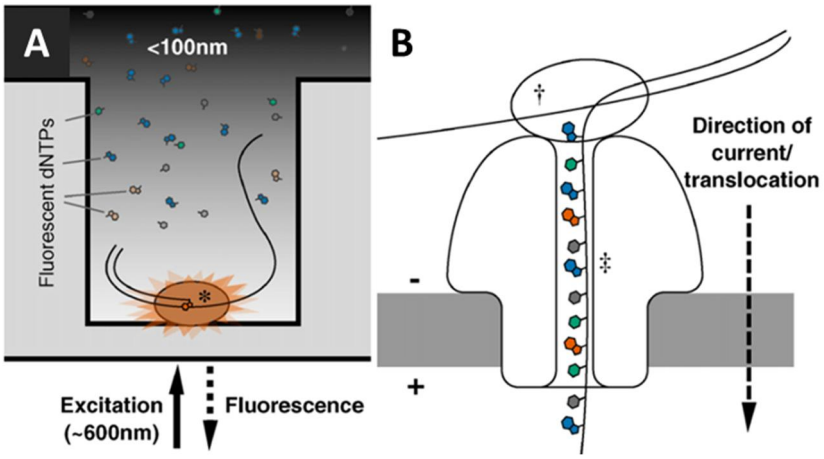


Figure 1 Single molecule sequencing nucleotide detection. (A) Nucleotide detection in a zero-mode waveguide (ZMW), as featured in PacBio sequencers. DNA polymerase molecules are attached to the bottom of each ZMW (*), and target DNA and fluorescent nucleotides are added. As the diameter is narrower than the excitation light's wavelength, illumination rapidly decays travelling up the ZMW: nucleotides being incorporated during polymerisation at the base of the ZMW provide real-time bursts of fluorescent signal, without undue interference from other labelled dNTPs in solution. (B) Nanopore DNA sequencing as employed in ONT's MinION sequencer. Double stranded DNA gets denatured by a processive enzyme (†) which ratchets one of the strands through a biological nanopore (‡) embedded in a synthetic membrane, across which a voltage is applied. As the ssDNA passes through the nanopore the different bases prevent ionic flow in a distinctive manner, allowing the sequence of the molecule to be inferred by monitoring the current at each channel [Heather and Chain, 2016].

PacBio SMRT technology (previously known as Nanofluidics) is based on the sequencing-by-synthesis method and optical detection of fluorescently labeled nucleotides (Figure 1A). This system is considered to be the most established [Roberts et al., 2013] and has already had several highly positive outcomes. However, it is worth noting that per-nucleotide accuracy (10-15% raw error rate) should still be improved. Despite obvious limitations, several projects have successfully exploited PacBio sequencing for high quality assemblies of large genomes [Berlin et al., 2015].

A more recent TGS technology, Oxford Nanopore, relies on measuring changes in electrical properties of DNA molecules when they pass through a pore (Figure 1B). Base calling of DNA is thus performed by detection of minute disruptions to electric current [Levy and Myers, 2016]. An obvious advantage of Oxford Nanopore platform is a handheld device MinION. Due to a low cost and a small size of the instrument, it was used for studies in very remote locations [Quick et al., 2016]. Unlike PacBio, Nanopore sequencing suffers from worse accuracy and a lower throughput.

Third-generation mapping technologies have been developed to complement sequencing data for the large-scale genome structure analysis by eliminating the need to sequence every base. At present, optical mapping is one of the most successful technologies on the market. In 2010 the technology was commercialized by BioNano Genomics and became available as a high-throughput platform termed Irys. The principle of Irys system is to fingerprint long DNA molecules by imaging the patterns of restriction sites under light microscopes using fluorescently labeled enzymes [Schwartz et al., 1993; Lam et al., 2012]. After imaging, such individual fingerprints can be assembled into larger optical maps, typically spanning many megabases of a chromosome [Valouev et al., 2006]. Although several studies demonstrated significant improvement in scaffolding, if combined with second-generation sequencing, optical mapping itself still suffers from biases. For example, incomplete nicking of the DNA produces a proportion of unlabeled digest sites during restriction, whereas multiple nick sites in close proximity to each other cause the DNA to shear, which in its turn limits the overall length of the map [Pendleton et al., 2015].

The other two recent third-generation mapping protocols are the cHiCago and GemCode. The first technology is based on Hi-C proximity ligation approach (described in detail below) and is proprietary to Dovetail. The cHiCago protocol captures chromatin interactions after confounding biological signals are removed by reconstructing *in vitro* chromatin [Putnam et al., 2016]. A need to ship the sample as well as to process it on site possibly limit the potential application of the cHiCago protocol.

The second technology, the Chromium instrument from 10X Genomics, is conceptually similar to the Illumina TruSeq Synthetic Long Read approach (reviewed in the appropriate section in the context of haplotyping as SLRH). In contrast to Illumina, 10X Genomics uses oil emulsion and multiple displacement amplification to amplify and ligate the barcodes across much longer molecules [Zheng et al., 2016]. Short reads, produced by Chromium, however, cannot be assembled into ‘synthetic’ long reads due to the very low sequencing coverage and are instead used for scaffolding [Mostovoy et al., 2016]. Therefore, 10X

Genomics data relies on the presence of a prior sequence assembly which itself decreases the power of the technology.

A recent example of a successful combination of third-generation sequencing strategies is a *de novo* assembly and haplotype phasing of the Korean individual [Seo et al., 2016]. The work by Chaisson et al., 2015 – demonstrates that long range sequencing can close a substantial number of existing sequencing gaps.

Despite a growing potential of new third-generation approaches, there are still several technical challenges which should be overcome until these technologies supersede the conventional sequencing performance of second-generation technologies: (i) very high per-base sequencing cost, (ii) quality in terms of per-base accuracy, (iii) accessibility of sequencing machines for small- and middle-scale research institutions, and, finally, (iv) contiguity of haplotype assemblies, i.e. longer reads should also be assembled into haplotypes.

Experimental approaches to link short reads

In the past ten years, a great number of original research articles introducing new methods, capable of accurate analysis of diploid genomes, have been published. This proves a growing interest to this hottest topic which importance was emphasized by several human genome-related initiatives, such as the 1000 Genomes Project and the International HapMap Project. Recently, a great variety of solid technologies for haplotyping was comprehensively reviewed by several groups [Snyder et al., 2015; Tu et al. 2016]. Although the authors define the technologies as technologies for haplotyping, this term unintentionally narrow the scope of these methods. All procedures, which are suitable for haplotyping, enable to solve a more complex task of restoring an uninterrupted genomic sequence. In this respect it might be more correct to use the term ‘contiguity preserving sequencing’. In the context of the further discussion these two terms are used as interchangeable.

Subselection of a particular part of a genome

Several studies have endeavoured to utilize chromosome separation strategy for phasing the entire human genome. For example, *Ma et al*. determined molecular haplotypes by laser capture microdissection [Ma et al., 2010]. The authors outlined the procedure, called 7 dimensional DNA (7DDNA), which entails parallel genotyping of both – a 7DDNA sample, containing only a subset of chromosomes, and an entire genomic sample. Upon determination of SNPs in each sample, a 7DDNA with homozygous genotype at all polymorphic loci along

a chromosome known to be heterozygotes would indicate that this sample contains one single copy of this chromosome. Microdissection involves cell fixation in metaphase, spreading the chromosomes into a microscopic slide and their subsequent isolation. After microdissection, subsets of chromosomes are subjected to multiple strand displacement amplification (MDA) and conventionally genotyped (Figure 2A).

In 2011 another group suggested using fluorescence-activated cell sorting (FACS) technology to separate chromosomes in lymphocytes [Yang et al., 2011]. They resolved most of the chromosomes by bivariate sorting with Chromomycin (binds guanine-cytosine-rich regions) and Hoechst (binds adenine-thymine-rich regions) staining (Figure 2B). Four unresolved chromosomes (Chr 9, 10, 11, and 12) with similar bivariate distribution patterns were identified by molecular typing which also served as quality control. The sorted single chromosomes were placed into wells of a 96-well plate, then amplified with MDA. During amplification, each copy of the chromosome was tagged by a short stretch of nucleotides to enable pooling and multiplexed sequencing. The authors are sure that, with respect to heterozygous SNP discovery, their approach, termed Phase-Seq, is supposed to be more reliable in comparison with the conventional genome sequencing.

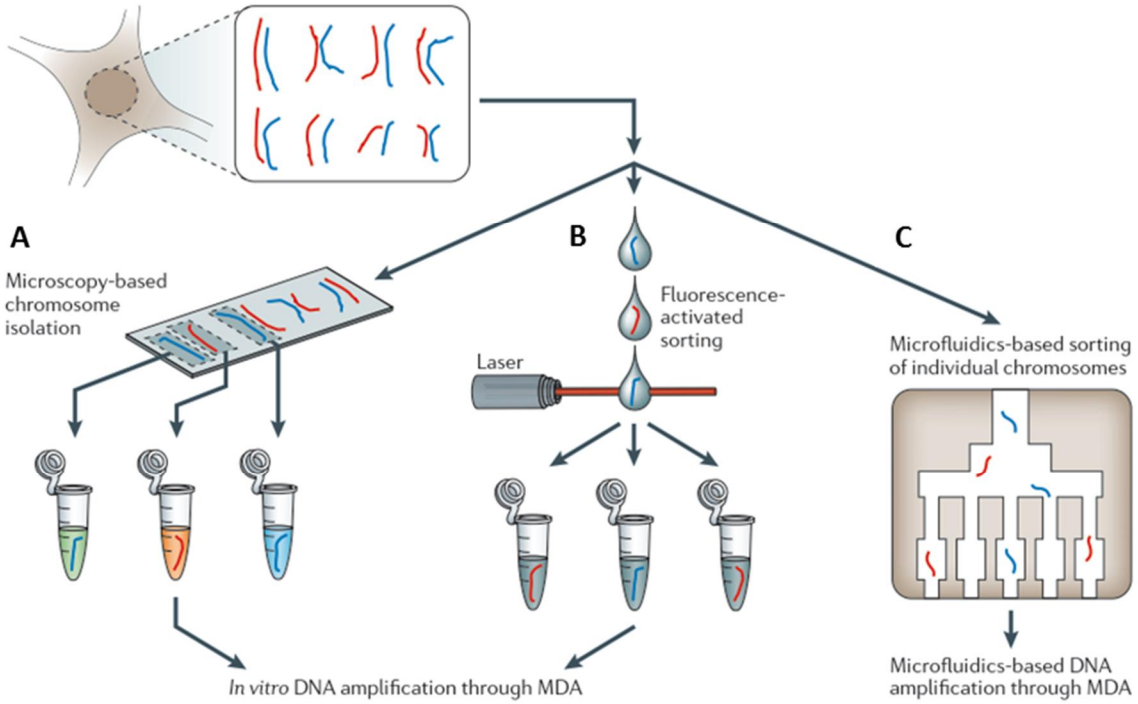


Figure 2 Schematic representation of the methods based on physical separation of chromosomes. Intact metaphase chromosomes from a single nucleus are isolated and compartmentalized by one of several means: (A) microscopy-based chromosome isolation (7DDNA), (B) fluorescence-activated sorting (Phase-Seq)**,** (C) microfluidics-based sorting (DDP). After compartmentalization by one of these methods, high-gain whole-genome amplification, such as multiple displacement amplification (MDA), is performed in each reaction chamber. Then, sequencing libraries are prepared from each amplified chromosome (not shown) [Snyder et al., 2015].

A direct deterministic phasing (DDP) method described by *Fan et al*. uses a microfluidic device to isolate individual chromosomes (Figure 2C) [Fan et al., 2011]. The instrument includes several regions designed to perform the following functions. The cell-sorting region is responsible for microscopical identification and the capture of single metaphase cells. The chromosomes are then released by protease digestion of the cytoplasm and randomly compartmentalized into 48 partitions in the chromosome release and chromosome partition regions correspondingly. Subsequently, the isolated chromosomes are individually amplified using MDA in the amplification region, followed by collection of the amplified products in the product retrieval region. The final step in the workflow of the DDP method is DNA genotyping either by standard single-nucleotide polymorphism arrays or massive parallel sequencing to produce haplotypes spanning an entire genome. Although the DDP approach is scalable, metaphase cells should be identified manually, therefore this procedure is considered as the most labor-intensive. The authors successfully validated their technique by phasing the genomes of the mother-father-child trio from the HapMap project. Moreover, they also

analyzed the fourth individual, whose genome had been already sequenced at highly polymorphic human leukocyte antigen (HLA) locus, confirming the potential of DDP for clinical diagnostics.

The common limitation for all of the above methods is the requirement for intact mitotic cells. In comparison with other direct haplotyping methods, microdissection is time-consuming and expensive. In the case of DDP or FACS sorting approaches, it is the dependence on sophisticated devices that can delay their routine use.

Notably, a classic strategy for determination of long-range haplotypes, such as somatic cell hybrids approach (conversion), in which single copies of one or a few human chromosomes are present within a fused mouse-human cell line, involves considerable cost and time, which is prohibitive for a large-scale application [Yan et al., 2000; Douglas et al., 2001; Marchini et al., 2006]. A combination of the conversion approach with the polony (polymerase colony) technology was reported in 2006. *Zhang et al.* immobilized diluted chromosomes within a polyacrylamide gel and subsequently genotyped them using serial PCR and single-base extensions [Zhang et al., 2006]. However, their approach was limited by the small number of heterozygous loci [Liu et al., 2008]. The efficient use of cell hybrids is also limited by the need for specialized and expensive equipment.

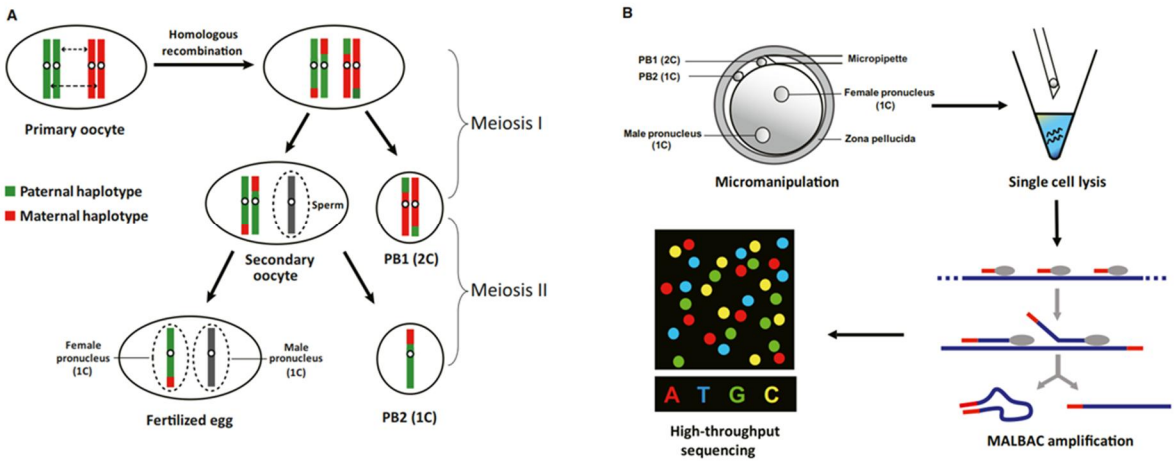


Figure 3 Schematic Charts of Human Oocyte Meiosis and Single-Oocyte MALBAC Sequencing (A) Illustration of homologous recombination and chromosome segregation during meiosis process of human oocytes. Only one chromosome is shown with red and green colors, indicating the maternal haplotype and paternal haplotype, respectively. (B) Flowchart of experiment procedures. The first and second polar bodies, dispensable for embryo development, were safely biopsied by a micropipette, followed by single-cell lysis, MALBAC amplification, and high-throughput sequencing. [Hou et al., 2013].

An alternative to physical separation of chromosomes is to leverage human gametes which have natural packaging of haploid complements. Recently, several studies have sequenced isolated sperm cell genomes using MDA or multiple annealing and looping-based

amplification cycles (MALBAC) [Wang et al., 2012; Lu et al., 2012; Kirkness et al., 2013]. Analogous to other reports, *Hou et al.* applied whole-genome amplification (MALBAC) to analyze the genome of single human oocytes [Hou et al., 2013]. The authors sequenced the triads of the first and the second polar bodies and the oocyte pronuclei from the same female egg donors. Using identified SNP's, they resolved the genomes of these donors and determined the crossover maps of their oocytes. Schematic charts of human oocyte meiosis and single-oocyte MALBAC sequencing are shown in Figure 3. Interestingly, in comparison with sperm, the oocytes exhibited higher recombination and aneuploidy rates. Moreover, the genome coverage with a comparable sequencing depth for a single oocyte was 20% higher than that demonstrated for a single sperm cell.

In principle, analysis of gametes can potentially yield complete haplotypes at chromosome-scale. However, due to a limiting amount of biological material, these methods heavily suffer from biased amplification. Usage of random primers in MDA or MALBAC can produce false variants through the formation of chimeric sequences or result in under-representation of GC-rich regions [Peters et al., 2012]. Also, the necessary tissues are not always readily available in non-clinical studies.

*Random subselection of **a** part of **a** genome*

Cloning pools

The methods which include physical chromosome separation during cell division require usage of complex specialized devices and careful manipulation of cells. This objective shortcoming of chromosome microdissections and related techniques motivated researchers to consider new possibilities for accurate analysis of the diploid human genome.

The framework underlying the first approaches for genome analysis by subselecting its random parts was first formulated in 1989 [Dear and Cook, 1989]. The concept was based on a physical linkage between markers on HMW DNA and relied on limiting dilution to subhaploid pools. Later, other groups extended the idea to clone-based approaches and exploited dilution pools for genome phasing. For example, *Burgtorf et al.* described a procedure for whole-genome haplotyping termed clone-based systematic haplotyping (CSH) [Burgtorf et al., 2003]. The main principle of CSH involves creating large-insert fosmid cloning and screening of the clones by PCR coupled with a mass spectrometry procedure for SNP typing. In 2011 *Kitzman et al.* suggested a similar but cost-effective approach for assembling long haplotypes using NGS platform for the identification of the variants [Kitzman et al., 2011]. A schematic representation of the procedure is shown in Figure 4. The

first step in their approach also included a generation of a fosmid library (~37kb insertions) which is then randomly partitioned into 115 pools, with each pool containing 3% of physical coverage of the diploid genome. The resulting pools were then used to prepare barcoded libraries with Nextera technology and, after combining, were shotgun sequenced on Illumina instrument. Overlaps between haplotypes derived from distinct pools were stitched together to assemble even longer haplotypes. More recently, *Lo et al*. exploited the same strategy for diploid assembly of a personal genome but, unlike the previous method, they used for cloning bacterial artificial chromosomes (BACs) with longer insertions (~140kb) [Lo et al., 2013]. The authors were able not only to achieve greater haplotypes blocks with N50 value of 1.6Mb but also provided the practical guidelines for the development and design of clone-based methods.

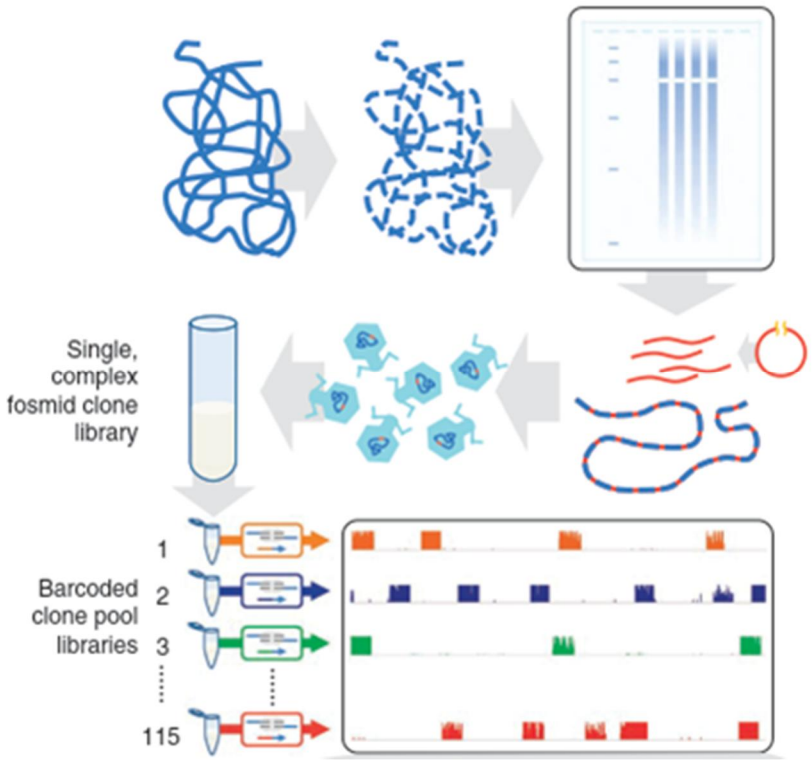


Figure 4 Haplotype-resolved genome sequencing. A single, highly complex fosmid library was constructed and split into 115 pools, each representing ~3% physical coverage of the diploid human genome. Barcoded shotgun libraries from each pool were constructed, then combined and sequenced [Kitzman et al., 2011].

Although library construction protocols are technically challenging, nevertheless at least 30 human genomes were resolved by fosmid or BAC clone dilution pools rather than by any other method [Snyder et al., 2015]. It is worth noting that construction of the human reference genome was performed by cloning of long fragments, which is similar to the cloning pools strategy [Venter et al., 2001; Lander et al., 2001].

The disadvantages of cloning pools technologies include: a large amount of initial genomic DNA, extensive library preparation protocols and, currently, a prohibitively high cost for use in a routine clinical environment. Furthermore, a sufficiently high number of *in vitro* pools with informative subhaploid content, from which indexed sequencing libraries are constructed, also limits possible applications of these methods. In addition, there exists an extra limitation to a size of the platform (BACs or fosmids) for the cloning-based methods [Levy et al., 2007; Lo et al., 2013].

Fragment pools

As was discussed above, a general dilution pools approach based on the fosmid or BAC libraries is highly labour- and resource-intensive. Despite the fact that dilution pools are capable of providing long-range genome analysis, one-by-one cloning and sequencing of many large fragments make this approach hardly scalable. The single-molecule dilution (SMD) concept [Jeffreys et al., 1990; Ruano et al., 1990], first described in early 1990s for studying multiple polymorphisms, is a simple technique but, due to a limited length of resulting PCR products, it is not suitable for determining haplotypes. Extraordinary advancements in and a variety of whole-genome amplification technologies made it possible not only to achieve significantly greater size of the preamplified fragments, but also to operate minute amounts of starting DNA resulting in sufficient yields and quality of the products for diverse downstream applications [Rykalina et al., 2014].

Several groups implemented dilution protocols relying on MDA instead of conventional amplification. In 2005 *Paul and Apgar* demonstrated a technique in which dilution of DNA to subhaploid equivalency followed by multiple strand displacement amplification was capable of separating di-allelic regions. The authors benchmarked their method by resolving a highly polymorphic HLA locus using two types of amplification. First, a haploid equivalent of a template DNA (3.5pg) was MDA amplified with about 10µg yield of the product. Second, the diluted product was subjected to the second amplification with HLA-specific primers. Assaying the sample on ABI Prism sequencer proved the applicability of the method for separating HLA-A, HLA-B and HLA-C alleles.

Seven years later, an updated version of the above mentioned approach, termed long fragment read (LFR) technology, was applied for genome-wide haplotyping [Peters et al., 2012]. Unlike single-molecule dilution with MDA, the LFR protocol utilizes DNA barcodes and Complete Genomics sequencing platform to analyze the NGS libraries. The principle of the fully automated method is shown in Figure 5. Briefly, 100-130pg of long parental DNA

fragments corresponding to 10-20 cell equivalent are stochastically compartmentalized into 384-well plate, amplified with MDA and through five enzymatic steps fragmented and ligated with barcoded adapters within each well. The resulting libraries are pooled, amplified with primers, common to the ligated adapters, and then analyzed using massively parallel short-read sequencing. The power of the LFR approach was validated on 7 genomes allowing to accurately resolve 84-97% of heterozygous variants.

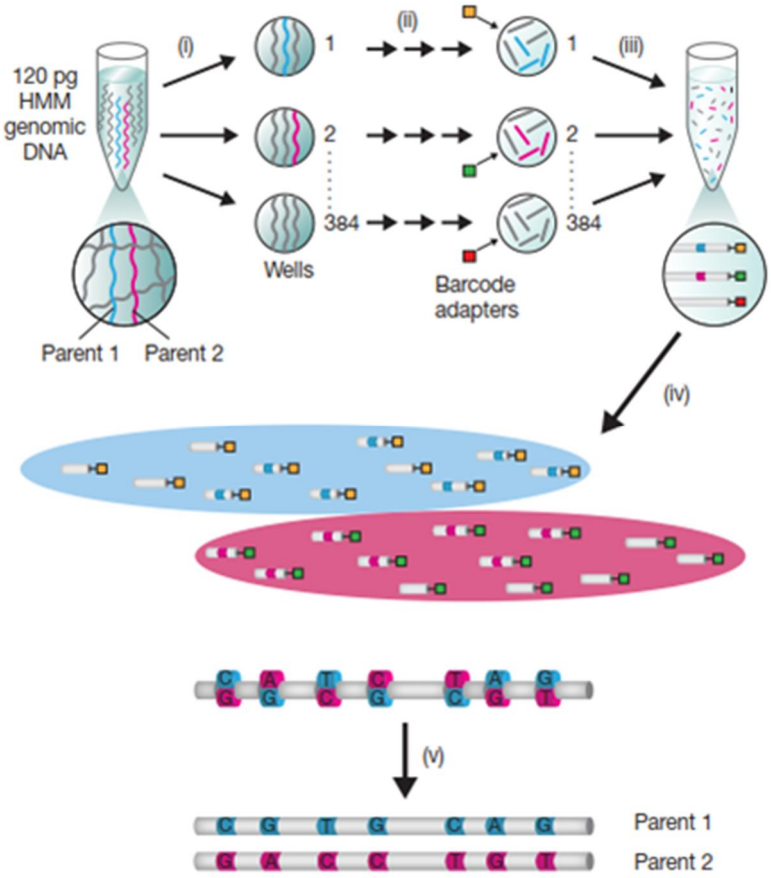


Figure 5 The long fragment read (LFR) technology. An overview of the LFR technology and controlled random enzymatic fragmenting is shown. (i) First, 100–130pg of high molecular mass (HMM) DNA is physically separated into 384 distinct wells; (ii) through several steps, all within the same well without intervening purifications, the genomic DNA is amplified, fragmented and ligated to unique barcode adapters; (iii) all 384 wells are combined, purified and introduced into the sequencing platform of Complete Genomics10; (iv) mate-paired reads are mapped to the genome using a custom alignment program and barcode sequences are used to group tags into haplotype contigs; and (v) the final result is a diploid genome sequence [Peters et al., 2012].

Likewise, in 2013 *Kaper et al*. successfully phased two human genomes (~95%), first demonstrating proof of concept by targeted haplotyping of the Duchenne muscular dystrophy region [Kaper et al., 2013]. In contrast to original LFR procedure they employed Nextera technology for barcoded library preparation and performed sequencing on Illumina platforms what made their approach accessible to any researcher.

More recently, a very similar approach, referred to as statistically aided haplotyping (SLRH), has been described by *Kuleshov et al*. [Kuleshov et al., 2014]. The SLRH procedure starts with DNA shearing and subsequent size selection in a gel to 8-10 kb fragments. Next, DNA fragments are ligated with amplification adapters and diluted into 384-well plate. The content of each well with 3000-6000 molecules is then PCR amplified using adapter-specific primers, followed by Nextera-mediated library preparation and barcoding through limited-cycle PCR. Finally, the resulting sublibraries are pooled down and analyzed on Illumina platform. The relatively short length of fragments in the initial step, compared with those obtained by related methods, was compensated by the development of Prism, a statistical phasing algorithm. The approach was validated by phasing 99% of single-nucleotide variants in three human genomes.

Notably, the LFR and SLRH technologies, based on fragment dilution methodology, are now commercialized by Complete Genomics and Illumina (Moleculo system) correspondingly.

Although fragment pools methods are considered to be rapid and cost-effective for genome phasing in comparison with other technologies, starting from minute amounts of DNA per pool can complicate accuracy, reproducibility and uniformity of amplification.

CPT-seq

In 2014 a novel approach for whole-genome sequencing termed contiguity-preserving transposition (CPT-seq) was described [Amini et al., 2014]. The approach was found to be a powerful tool for both haplotype-resolved sequencing and later for use in *de novo* assembly scaffolding by means of a specially designed fragScaff algorithm [Adey et al., 2014].

Hyperactive Tn5 transposomes were successfully used for preparation of NGS libraries due to their ability to simultaneously fragment DNA and append adapters at about 300 intervals, which was referred to as tagmentation [Adey et al., 2010]. Besides creating DNA libraries in one enzymatic step, Tn5 transposase possesses another unique property, it stays bound to target DNA molecules after tagmentation until an agent responsible for dissociation of DNA-protein complex is added to the mixture.

The method overview is shown in Figure 6. The CPT-seq primarily utilizes inherent property of Tn5 transposomes to incorporate indexed adapters while physically holding adjacent library molecules. Tagmentation is carried out on 1ng (about 300 haploid human genomes) of high molecule weight DNA samples tagging the sequences with a unique combination of adapters, resulting in total of 96 indices. These reactions provide an initial indexing tier. After this step, a subhaploid dilution and compartmentalization are performed. The content of the

96-well plate is pooled down, appropriately diluted and split into new pools. Within this additional subsequent assortment of the molecules, the Tn5 protein is released from fragmented DNA templates which are then amplified in 96 indexed PCR reactions thus producing the second index tier with 9216 virtual compartments. The approach leverages minimal hands on time (3 hours) and readily available equipment. More recently, a detailed version of the protocol has been published with an application for single cell ATAC-seq [Christiansen et al., 2017].

Overall, there are two technological advances underlying the method: (i) contiguity-preserved indexing via tagmentation and (ii) combinatorial split-and-mix strategy generating about 10000 distinct virtual compartments. To note, a relatively high amount of DNA per physical compartment (3 copies of the genome) allows robust PCR, whereas virtual compartments with only 3% of haploid content facilitate avoidance of collision of parental DNA fragments. The latter dramatically lessens the dilution factor needed to reach subhaploid scale (content), if compared with similar technologies.

It is also worth mentioning, the authors themselves have recently emphasized several shortcomings of their CPT-seq protocol. Although the overall amount of input DNA required is high (nanograms rather than picograms), only a small portion of the library is sequenced in each partition [Snyder et al., 2015]. Moreover, a subhaploid portion of molecules is sequenced with low coverage due to a 50% loss during amplification owing to two different adapters (A and B) for tagmentation. Molecules with AB adapters only are amplified in PCR. In addition, a broad range of library size provides advantage in PCR for those with shorter insertions [Amini et al., 2014].

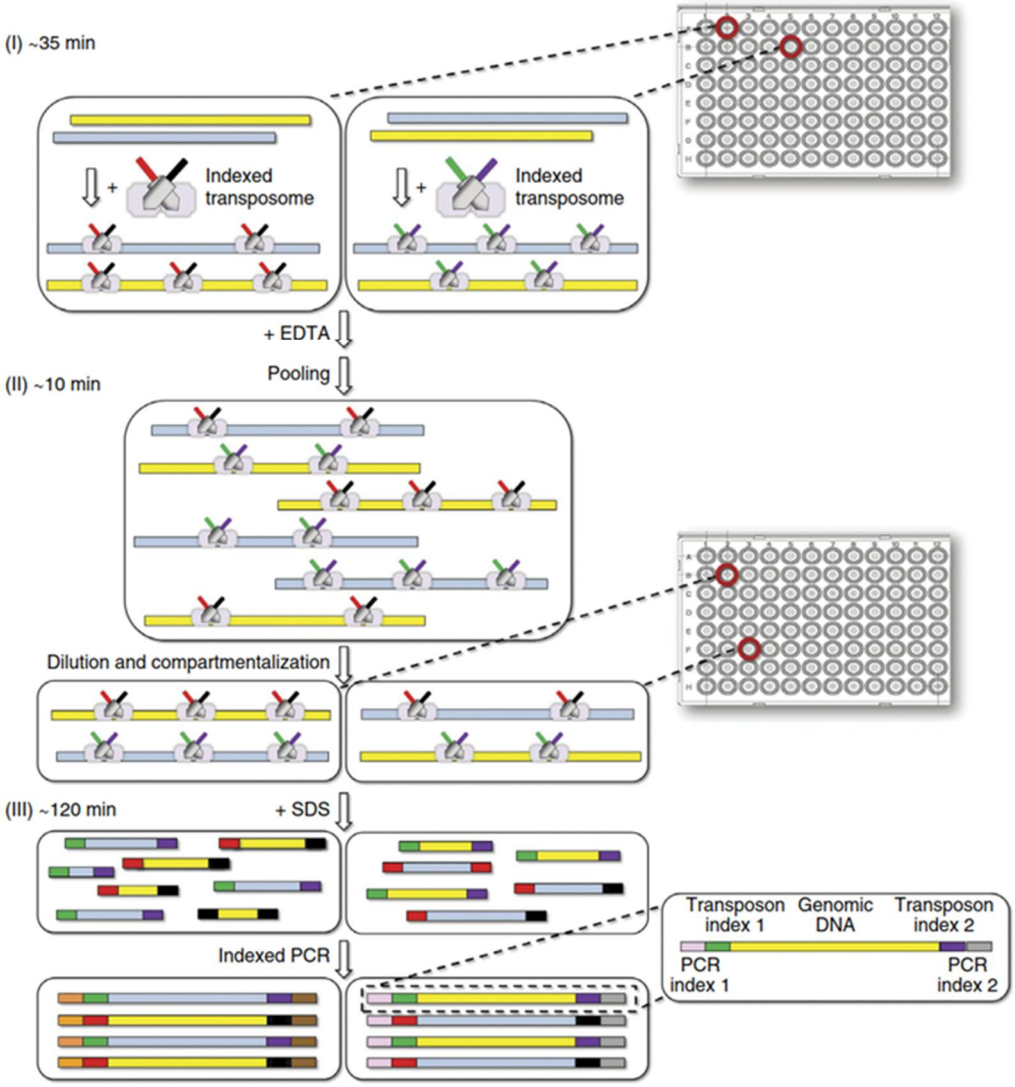


Figure 6 Overview of the CPT-seq workflow. There are three key steps: (I) indexed transposition, (II) pooling, dilution and compartmentalization, and (III) indexed PCR. A set of 96 different indexed transposome complexes are used to set up 96 independent transposition reactions to create separate virtual genomic partitions (step I). Transposition reactions are pooled together, diluted to subhaploid DNA content and split into 96 compartments (step II). Upon removal of the transposase with SDS, compartment-specific libraries are generated using indexed PCR (step III). All samples are pooled together after PCR and prepared for sequencing [Amini et al., 2014].

From our point of view, the resulting price for the reagents might be the main limitation of CPT-seq method. Although the authors apply a combination of only 20 (8+12) oligonucleotides to create the 96 asymmetrically indexed transposomes and the same amount for indexed PCR, the overall number of oligonucleotides is still high. More importantly, it could be quite expensive to assemble 96 transposome batches (with 2.5 pmol for a single tagmentation) using Epicentre Tn5 transposase or just impossible with already preassembled Illumina transposomes. However, the protocol cost can drop down, if in-house Tn5 protein is adapted to the protocol.

Proximity ligation strategies

Crosslinking and proximity ligation approach was originally developed to provide detailed information about a three-dimensional folding of a genome. The idea behind this strategy was to link physically interacting regions in chromatin by means of spatially constrained ligation. The first method that followed this principle was chromosome conformation capture (3C), described in 2002 [Dekker et al., 2002]. The 3C technique was initially validated on yeast nuclei and, to analyze long-range interactions within and between chromosomes, used locus-specific amplification. Adaptation of 3C evolved in chromosome conformation capture-on-chip (4C), a version with inverse PCR or 3C-carbon-copy (5C) which employed highly multiplexed ligation-mediated amplification [Simonis et al., 2006; Dostie et al., 2006].

Further development of the approach made use of coupling proximity-based ligation with massively parallel sequencing (Hi-C) [Lieberman-Aiden et al., 2009]. The main advantage of Hi-C in comparison with its precursory methods is that it does not require the selection of a set of target loci prior to analysis, while permitting to ligate two distinct regions of a chromosome in a single sequencing read.

Schematic representation of Hi-C is outlined in Figure 7. Intact cells or nuclei are subjected to formaldehyde fixation to crosslink proteins with DNA and with other proteins. This step allows to connect DNA sequence segments which were close to one another in the nucleus. After crosslinking, DNA is cleaved with a restriction enzyme, filled-in at 5`-overhangs with biotinylated nucleotides followed by intramolecular ligation of blunt-ended fragments under dilute conditions. Next, the ligated DNA fragments with biotin at the junction are sheared and purified using streptavidin beads. The resulting Hi-C library is then amplified and sequenced generating a catalog of interacting fragments.

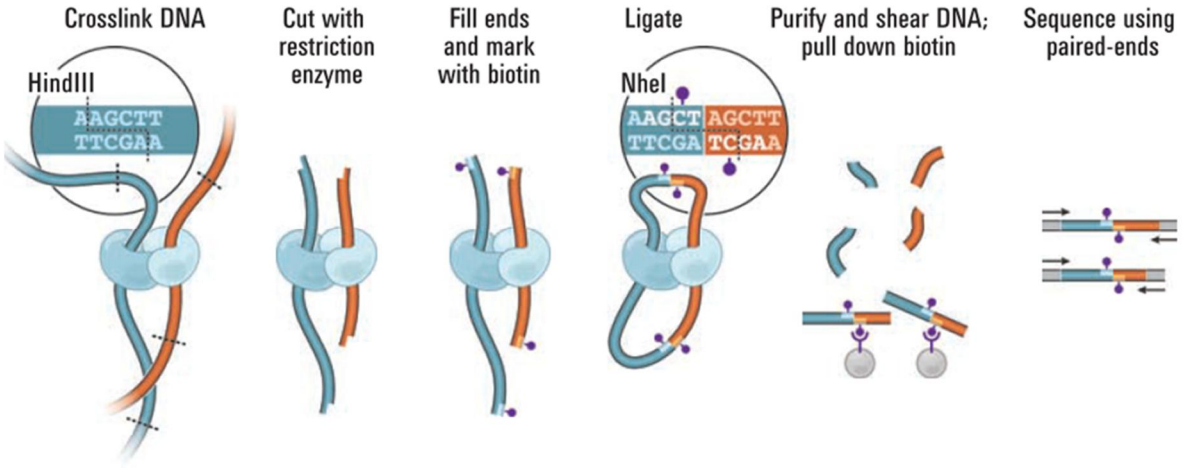


Figure 7 Overview of Hi-C. Cells are cross-linked with formaldehyde, resulting in covalent links between spatially adjacent chromatin segments (DNA fragments shown in dark blue, red; proteins, which can mediate such interactions, are shown in light blue and cyan). Chromatin is digested with a restriction enzyme (here, HindIII; restriction site marked by dashed line; see inset), and the resulting sticky mends are filled in with nucleotides, one of which is biotinylated (purple dot). Ligation is performed under extremely dilute conditions to create chimeric molecules; the HindIII site is lost and an NheI site is created (inset). DNA is purified and sheared. Biotinylated junctions are isolated with streptavidin beads and identified by paired-end sequencing [Lieberman-Aiden et al., 2009].

Several groups have recently pioneered application of Hi-C probability maps to produce chromosome-scale assemblies and haplotypes using short-read data [Burton et al., 2013; Kaplan et al., 2013; Selvaraj et al., 2013; de Vree et al., 2014]. For instance, *Burton et al.* developed an algorithm termed LACHESIS which generated *de novo* assemblies for human, mouse and fruit fly genomes. Their computational approach exploits combination of Hi-C interactions with shotgun fragments and long-range sequencing data. A further extension of LACHESIS was demonstrated by *Kaplan et al.* who predicted the location of 65 unplaced contigs in the human genome. Likewise, *Selvaraj et al.* were first to apply Hi-C data to phase SNPs onto haplotypes at chromosome-scale. The authors called their method HaploSeq which was successfully validated both on a hybrid mouse embryonic stem cell line and human lymphoblastoid cell line. Finally, *de Vree et al*. based on a conceptually similar principle, selectively sequenced and phased entire genes.

Although crosslinking and proximity ligation technologies proved to be invaluable for long-range genome analysis, they still have several limitations to be considered. First, Hi-C signals are better at extracting information about the order of chromosomal segments rather than information about their orientation [Korbel and Lee, 2013]. Second, performance of Hi-C approaches depends heavily on the length of DNA reads because shorter reads are challenging

to map into multiple locations in a genome and, thus, to obtain optimal level of contiguity [Adey et al., 2014]. Finally, Hi-C and related methods still require a large amount of intact cells or nuclei which is not always available. However, a recent work of *Putnam et al*. where chromatin was reconstituted *in vitro*, shows great promise that the latter limitation of proximity-ligation-based technologies might be eliminated in the future [Putnam et al., 2016].

Paired indexing of fragmentation sites

The strategies described above aim at revealing fragments originating from the same initial DNA molecule or a group of molecules. We thought it would be feasible to look at the problem from a different angle and focus on fragmentation sites. After all, the problem in genome reconstruction is that fragmentation sites cannot be unambiguously mapped. If it were possible to determine which two fragment ends arise from the same fragmentation site, then the correct order of the fragments could be restored and the original molecule sequence determined. Figure 8 illustrates this idea in comparison to existing approaches using the example of haplotype determination task. In a standard NGS library (variant **1**) the fragments lose any indication of their chromosome of origin; after sequencing it is not possible to determine which combinations of detected allelic variants are present in the input material, (e.g. fragments *a-d-f* or *a-j-l*). When independent libraries are prepared from subhaplotype amounts of the input material (variant **2**) – sequenced fragments originating from the same molecule can be grouped. In the proposed approach (variant **3**) fragmentation sites are marked, so that after sequencing it can be determined that fragments *j* and *k* were adjacent in the molecule of origin. The more fragments get sequenced the longer scaffolds would be assembled. Labeling fragmentation sites would resolve duplications – fragments *e* and *f* would be revealed to be adjacent and not overlapping, as might happen in variant **1** and **2**. It should be stressed that labeling fragmentation sites makes the sequence assembly process independent from an internal sequencing context of the fragments. Even if several adjacent fragments consist exclusively of short tandem repeats – they would be assembled correctly, because external fragmentation site labels, rather than repetitive internal sequences, would be used as an assembly clue.

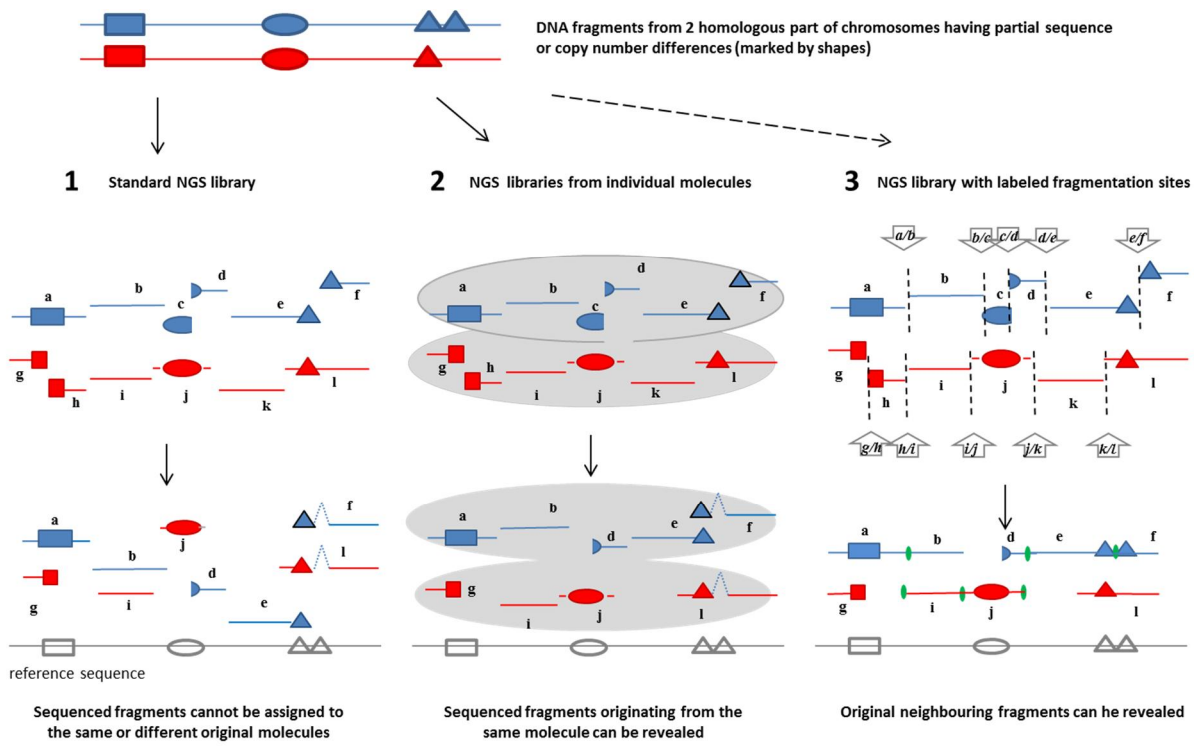


Figure 8 Principle of fragmentation sites labeling approach on the example of haplotype reconstruction task. During standard sequencing library preparation fragments linkage is completely lost (1). Separation of homologous molecules prior to fragmentation allows to group sequences belonging to one haplotype (2), grouped fragments are within grey ovals). Labeling fragmentation sites approach increases the degree of fragments grouping accuracy to the neighbouring fragments (3), arrows with letters show labeled fragmentation sites).

Practically, fragmentation sites can be labelled by attaching index sequences to the ends of the fragments arising at those sites. Comparatively short nucleotide stretches provide a huge variety of indices and are widely used to label nucleic acid molecules in molecular biology. In the NGS field unique molecular identifiers are attached to library molecules before amplification so it is possible to distinguish between independent library molecules and PCR duplicates [Kinde et al. 2011]. In the case of fragmentation site labelling, paired indices are required to be attached to the two fragment ends. Indices in the pair may be identical, or complementary to each other. They may also be two completely different stretches of sequences; it should only be known in advance which pairs of sequences can label a fragmentation site. After indexing, each fragment (except for terminal fragments) would bear two indices at its end, corresponding to two fragmentation sites. These indices could then be processed as part of the fragments or sequencing adapters, sequenced and used to join the ends of the fragments. A schematic overview of paired indexing of fragmentation sites approach is shown in Figure 9. In the ideal case, each fragmentation site is labeled and the assembly is univocal.

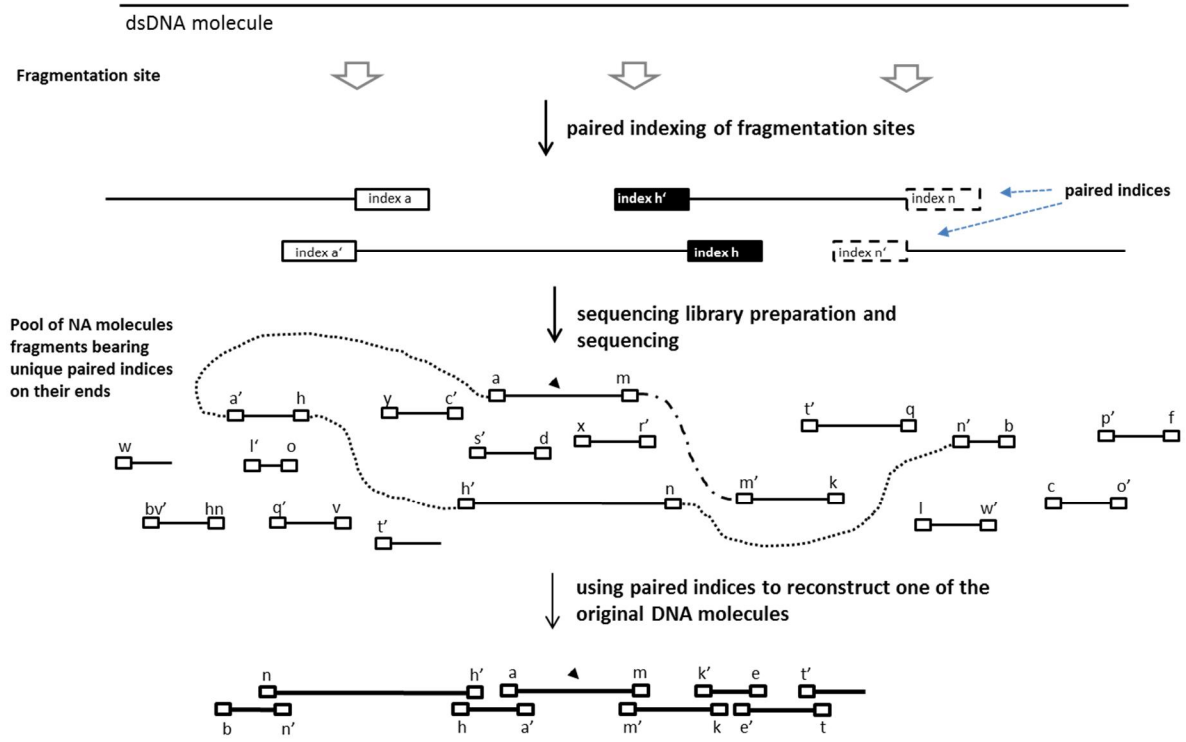


Figure 9 Paired indexing of fragmentation sites. Using different paired indices (shown as rectangular blocks with different fillings and different names) allows to mark individual fragmentation sites. Paired nature of indices preserves information about the neighboring fragments in the original DNA molecule. Algorithm for reconstruction is fairly easy: for an arbitrary starting fragment (here *a-m*, marked with an arrow), neighboring fragments bearing the paired indices a´ and m´ are determined for both sides, and so on - the indices on the free ends of the growing chain determine the next fragments to be added to the assembly. The search process is visualized here by two dashed lines beginning at the two ends of the starting fragment.

Attachment of the indices to the fragments should occur after fragmentation, which provides free DNA ends, but before physical separation of the fragments. Fragmentation using Tn5 transposase naturally fulfills these requirements. Tn5 transposase inserts symmetrical breaks into each strand of dsDNA and ligates transposon sequences to the 5´ends of the arising fragments, - this reaction is called tagmentation. The active unit in tagmentation is not the transposase itself but a transposome - a Tn5 dimer where each transposase molecule is bound to a specific 19nt double stranded transposon end sequence (mosaic end - ME sequences). ME sequences are obligatory for transposomes formation. Additional sequences – e.g. technical regions for a sequencing platform - may be introduced within an uninterrupted sequence in between the two ME regions (full transposon) or through an extension of a single ME region (transposon end adapter). Figure 10 shows possible locations of paired indices parts for both variants. In the case of transposon end adapters, indexed fragments get separated after removal of the transposase (Figure 10A). In the case of full transposons, the fragments would

still be held together by the double-stranded region of the transposon, therefore an additional fragmentation site (e.g. destroyable nucleotide) is required (Figure 10B).

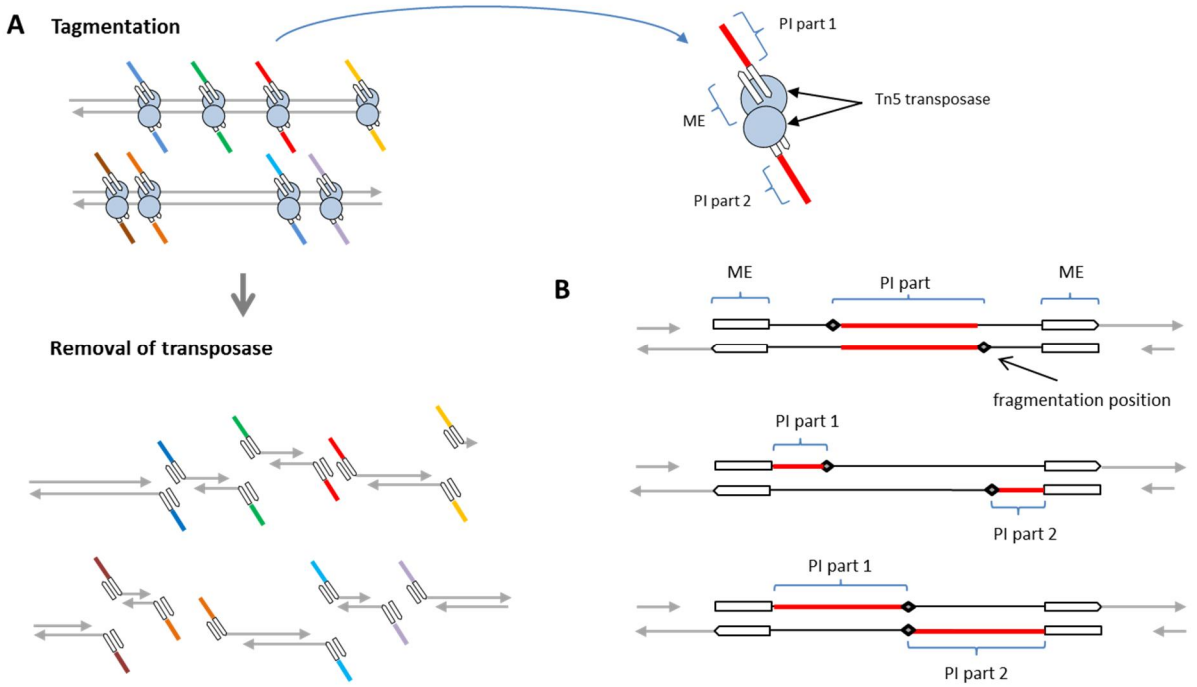


Figure 10 Examples of transposon structures bearing paired indices. (A) Tagmentation with transposon end adapters. Transposomes are depicted as double circles, each circle with partly double arrows, corresponding to transposase dimers bound to transposon end sequences. Transposase recognition sites (ME) are shown as empty double arrows. Paired indices (PI) are within single-stranded tails. Different colors represent unique index pairs in each transposome. After removal of transposase indexed fragments diverge in solution. (B) Full transposons inserted into the dsDNA hold the DNA fragments together, even after removal of transposase. Fragmentation positions are the positions where fragmentation following transposition occurs. Location of fragmentation positions and PI parts are designed in a way to make insertion of indices independent from the transposon orientation.

Transposomes are preassembled from transposase and oligonucleotides before tagmentation and are very stable. This feature makes it possible to prepare transposomes with paired indices independently from tagmentation reaction and keep the contents of the index pairs under control.

We independently came to the idea of using paired indices to label DNA ends arising at fragmentation sites and using those codes to link neighbouring fragments post-sequencing. However, when we were preparing the patent application, similar patents appeared in open access databases. Patent PCT/US2012/023679 basically claims the general idea of pairwise labeling of tagmentation fragments at fragmentation sites and suggests several possible transposons structures. PCT/US2011/059642 protects various modifications of an artificial transposon structure suitable for paired indexing at tagmentation sites. PCT/US2015/038050 describes a variation of tagmentation reactions where just one strand of dsDNA is tagmented

– one-sided transposition. This is an interesting approach because, even after tagmentation with transposon end adapters, the fragments are still kept together. PCT/US2014/065491 suggests a non-transposase based method of inserting paired indices as a part of loop structures containing random sequences which hybridize to the single-stranded DNA molecules.

Though the above mentioned patent applications describe the principle of the approach and suggest multiple ways of realization, they do not provide any real experimental data to support their suggestions and rather reflect the desire of the inventors to most broadly protect the idea in theory. Our group has been working in the field of NGS related methodology and technology development for a long time [Parkhomchuk et al., 2009; Borodina et al., 2011; Rykalina et al., 2014]. Solving the problem of the loss of sequence contiguity information in the course of NGS library preparation has always been in the focus of our interests. Paired indexing of fragmentation sites approach looks especially attractive for reconstruction of repetitive regions and distinguishing homologous sequences irrespective of the degree of their similarity. If established, this method would be of an extreme value both for resequencing and sequencing *de novo* applications. It was therefore decided to explore the PITS approach experimentally.

MATERIALS AND METHODS

Materials

Chemicals

Name	Company
Agar (bacteriology grade)	BD
Agencourt AMPure XP Beads	Beckman Coulter
Ammonium Persulfate (PSA)	Sigma
Ampicillin Sodium Salt	Sigma
ATP, adenosine 5’-triphosphate (100 mM)	GE Healthcare
Bacto-tryptone	Roth
Chitin Magnetic Beads 25ml	New England Biolabs
cOmplete, EDTA-free	Roche
Coomassie Brilliant Blue	Bio-Rad
dNTP Set (100mM of each A,C,T,G)	GE Healthcare
DTT 1,4-dithiothreitol (1M stock solution)	Sigma
EDTA disodium salt dihydrate (0.5M stock solution)	Ambion
Ethanol (absolute)	Merk
Formamide	Sigma
GelRed Nucleic Acid Gel Stain (10,000X in water)	Biotium
Glycerol	Merk
HEPES pH 7.2 (1M stock solution)	Sigma
Human Genomic DNA	Bioline
Hydrochloric Acid (HCl), 37%	Merk
HyperLadder 100bp	Bioline

Name	Company
HyperLadder 1kb	Bioline
IPTG isopropylthio-β-D-galactoside	Sigma
Lambda DNA	New England Biolabs
Mouse Genomic DNA	Bioline
Orange G	Sigma
PEI polyethyleneimine (50% stock solution)	Sigma
PlusOne Acrylamide PAGE	Amersham Biosciences
PlusOne Methylenbisacrylamide	Amersham Biosciences
PlusOne Urea	Amersham Biosciences
Polyethylene Glycol (PEG 6000)	Merk
Potassium Chloride (KCl)	Applichem
Precision Plus Protein Dual Color Standards	Bio-Rad
Ribonucleic Acid, transfer from baker's yeast	Sigma
RNaseZap RNase Decontamination Solution	Life Technologies
SDS sodium dodecylsulfate (20% stock solution)	Ambion
SOC Medium	New England Biolabs
Sodium Chloride (NaCl)	Sigma
Sodium hydroxide, pellets (NaOH)	Applichem
SYBR Green II Gel Stain (10,000x concentrate in DMSO)	Thermo Fisher Scientific
T7 Express lysY/Iq Competent E. coli	New England Biolabs
Triton X-100	Sigma
Tryptone BD	Roth
Tween-20	Sigma

Name	Company
UltraPure 10X TBE Buffer	Invitrogen
UltraPure Agarose	Invitrogen
UltraPure Low Melting Point Agarose	Thermo Fisher Scientific
UltraPure TEMED	Invitrogen
Yeast Extract BD	Roth
β-Mercaptoethanol	Merk
20/100 Oligo Ladder	IDT
2-Propanol	Merk
4x Laemmli Sample Buffer	Bio-Rad
6x Loading Dye Solution	New England Biolabs
5x Loading Dye Solution	Bioline

Hardware and Plastic

Name	Company
Agarose Gel Chambers	Home-Made
Alumina Plates for Vertical Slab Gels	Amersham Biosciences
Amicon Ultra-0.5 Centrifugal Filter Unit	Millipore
Amicon Ultra-4 Centrifugal Filter Unit	Millipore
Bioanalyzer 2100 Instruments	Agilent Technologies
Centrifugation tubes	Beckman
Centrifuge 5404	Eppendorf
Centrifuge 5415D	Eppendorf
Centrifuge 5810R	Eppendorf

Name	Company
Centrifuge Avanti J-26S	Beckmann Coulter
DNA Engine Thermal Cycler	Bio-Rad
DNA LoBind Tube (0.5ml)	Eppendorf
DNA LoBind Tube (1.5ml)	Eppendorf
DNA LoBind Tube (2.0ml)	Eppendorf
DNA LoBind Tube (5ml)	Eppendorf
Dual Gel Caster for Mini Vertical Units	Hoefer Inc.
DynaMag-15 Magnet	Thermo Fisher Scientific
DynaMag-2 Magnet	Thermo Fisher Scientific
Filter Tips	Axygen
Filters 0.22/0.45µm	Millipore
GE Healthcare GeneQuant 1300 Spectrophotometer	Thermo Fischer Scientific
Glass Plates for Vertical Slab Gels	Amersham Biosciences
Incubator IN110	Memmert
Incubator Shaker Innova 4430	New Brunswick Scientific
MicroAmp 8-Cap Strip	Applied Biosystems
MicroAmp 8-Tube Lids	Applied Biosystems
Microfluidizer Processor M-110 L	Microfluidics
Nanodrop 1000 Spectrophotometer	Thermo Fisher Scientific
Paraffin Paraplast Plus	VWR International
Parafilm M	Roth
Pasteur Pipets (plastic)	VWR International
Petri Dishes	Greiner Bio-One

Name	Company
Pipetboy Acu Integra	VWR International
Plastic Tubes (Polypropylene, 15 and 50 ml)	Greiner Bio-One
Qubit 2.0 Fluorometer	Life Technologies
Safe-Lock Tubes (0.5ml)	Eppendorf
Safe-Lock Tubes (1.5ml)	Eppendorf
Safe-Lock Tubes (2.0ml)	Eppendorf
SARSTEDT Serological pipets (5, 10 and 25 ml)	Greiner Bio-One
Single use Protective Gloves TNT (Nitrile, powder-free)	VWR International
Single-use Scalpels (Cutfix)	VWR International
StepOne Real-Time PCR System	Thermo Fischer Scientific
Sterile Needles BD Microlance 3 (21G, 25G)	Roth
Syringes BD Discardit II (1ml, 5ml and 10ml)	Roth
Thermomixer 5436	Eppendorf

Enzymes

Name	Company
Ampligase Thermostable DNA Ligase	Biozym
E. coli DNA Ligase	New England Biolabs
EcoRI	Invitrogen
Immolase DNA Polymerase	Bioline
T4 DNA Ligase	New England Biolabs
T4 DNA Polymerase	New England Biolabs
T4 Polynucleotide Kinase	New England Biolabs

Kits

Name	Company
EZ-Tn5 <KAN-2> Insertion Kit	Biozym
High Sensitivity DNA Analysis Kit	Agilent Technologies
MinElute PCR Purification Kit	Qiagen
Nextera DNA Library Preparation Kit	Illumina
Nextera Rapid Exome Kit	Illumina
QIAprep Spin Miniprep Kit	Qiagen
QIAquick Gel Extraction Kit	Qiagen
QIAquick PCR Purification Kit	Qiagen
Quant-iT dsDNA HS Assay Kit	Invitrogen
Qubit Protein Assay Kit	Thermo Fischer Scientific
SureSelect XT2 Reagent Kit	Agilent Technologies
SYBR Green PCR Core Reagents	Life Technologies
SYBR Select Master Mix	Thermo Fischer Scientific

Oligonucleotides

All oligonucleotides were purchased either from IDT, TIB or Sigma. The quality of the oligonucleotides was assessed by analysis on a denaturing polyacrylamide gel. The oligonucleotide concentrations were determined spectrophotometrically by measuring dilutions in 1X STE in accordance with the absorbance at 260 nm (A260) as described [Adams, 2003].

Oligonucleotides (5' → 3')	**Purpose**
T7 promoter (forward): TAATACGACTCACTATAGGG Seq_Rev_Tn5 (reverse): GATTGCCATGCCGGTCAAGG	Tn5 Production Sanger Sequencing
Tn5ME-A: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG Tn5ME-B: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG Tn5MErev: pCTGTCTCTTATACACATCT (p – phosphate) Underlined regions correspond to the double-stranded part of the adapter	FEA Transposon end adapters
p_015 (forward): GCTCACTCAAAGGCGGTAAT p_016 (reverse): GCTGGCGTAATAGCGAAGAG	FEA /2248bp PCR product
p_017 (forward): TTAGCAGAGCGAGGTATGTAGG p_018 (reverse): CATTTCCGTGTCGCCCTTATT	FEA /1240bp PCR product
p_019 (forward): CCTATCTCAGCGATCTGTCTATTTC p_020 (reverse): GCGCGGTATTATCCCGTATT	FEA / 610bp PCR product
p_021 (forward): CGGCTCCAGATTTATCAGCAATA p_022 (reverse): GCCAACTTACTTCTGACAACGA	FEA /310bp PCR product
p_015 (forward): GCTCACTCAAAGGCGGTAAT p_023 (reverse): CTTCAGCAGAGCGCAGATAC	FEA /597bp PCR product
p_025 (forward): CTTTCACCAGCGTTTCTGGG	FEA /602bp PCR product

Oligonucleotides (5' → 3')	**Purpose**
p_016 (reverse): GCTGGCGTAATAGCGAAGAG	
p_026: CTGTCTCTTATACAddC (bottom) p_027: CTGUCTCUTAUACAC (bottom) Tn5ME-A: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG (top) (ddC – dideoxycytidylate, U- uridine) Underlined regions correspond to the double-stranded part of the adapter	PITS Transposon end adapters
p_029: AGATGTGTATAAGAGACAGNNNNNNNNNATTACCGCCTTTGAGTGAGinvT In combination with oligonucleotides p_015 and p_028 (invT – inverted deoxythymidylate)	Gap Repair Model
p_028: pCTGTCTCTTATACACATCTCCGAGCCCACGAGACinvT (p – phosphate; invT – inverted deoxythymidylate)	PTULI Gap Repair replacement oligonucleotide
p_032 (forward): AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTC p_033 (reverse): CAAGCAGAAGACGGCATACGAGATGGATGTTCTGTCTCGTGGGCTCGG Underlined region correspond to the barcode	PTULI PCR

Oligonucleotides (5' → 3')	**Purpose**
p_008: CTGTCTCTTATACACATCTCCGAGCCCACGAGACUC(Biotin-dT)UUAATGATACGGCGACCACCGAGATCTACACCAGTTCGCGAAAAACGTG p_009: CTGTCTCTTATACACATCTCCGAGCCCACGAGACUCTUUAATGATACGGCGACCACCGAGATCTACACCACGCATTTTTCGCGAACT p_010: GTGTAGATCTATTCAATGAATCGGCCTACGTCGTCGGCAGCGTCAGATGTGTATAAGAGACAG (Biotin-dT – biotinylated nucleotide; U - uridine)	Alternative full transposon adapter: Lampion structure preparation
p_005 (forward): CTGTCTCTTATACACATCTTACGACCUGUGCAG(Biotin-dT)GTG p_006 (reverse): CTGTCTCTTATACACATCTACACCCAGTTTGGATTCTCC (Biotin-dT – biotinylated nucleotide; U - uridine)	Alternative full transposon adapter: 158bp PCR product
p_011 (forward): CTGTCTCTTATACACATCTCAGCCAGTGTGTCCCTTT p_012 (reverse): CTGTCTCTTATACACATCTGCGCCATCAAATGTGTGTAG	Alternative full transposon adapter: 258bp PCR product
p_013 (forward): CTGTCTCTTATACACATCTCCCAAGCTTCAGCCATTACT	Alternative full transposon adapter:

Oligonucleotides (5' → 3')	Purpose
p_014 (reverse): CTGTCTCTTATACACATCTCCTGCATCAGGCTTCTTCTT	495bp PCR product

Plasmids

Name	Company
pBluescript II KS (+)	Addgene
pTXB1-Tn5	Addgene
pUC19	New England Biolabs

Solutions

Name	Recipe
1000x Ampicillin (100mg/ml)	2g of sodium ampicillin salt was dissolved in 20ml H_2O. The solution was sterilized by filtration and stored at -20°C in 1ml aliquots.
1M IPTG	1g IPTG was dissolved in 3.5 ml H_2O then adjusted with up to 4.2ml. The solution was sterilized by filtration by passing it through a 0.22-µm disposable filter, dispensed into 520µl aliquots and stored at -20°C (under hood).
10% (w/v) Triton X-100	5g of 100% Triton X-100 was made up with H_2O to 50ml; mixed well and stored at RT.
4x Laemmli Sample Buffer	20µl β-mercaptoethanol was added to 180µl of Sample Buffer and mixed well (1:10).
cOmplete, EDTA-free Protease Inhibitor Cocktail	2 tablets of cOmplete were dissolved in 1ml of H_2O (1 tablet in 500µl) and stored at -20°C. The stock solution

Name	Recipe
(stock solution)	was stored at 2 to 8 °C for 1 to 2 weeks, or at least 12 weeks at -15 to -25 °C.
HEX Buffer (100ml)	18g NaCl was mixed with 2ml 1M HEPES-KOH (pH 7.2), 200µl 0.5M EDTA, 2ml 10% Triton X-100 and adjusted to 100ml with H_2O. The buffer was stored at at 4°C.
HEGX Buffer (100ml)	18g NaCl was mixed with 2ml 1M HEPES-KOH (pH 7.2), 200µl 0.5M EDTA, 2ml 10% Triton X-100, 10ml 100% Glycerol and adjusted to 100ml with H_2O. The buffer was stored at 4°C.
HEGX Buffer/cOmplete	800µl of Protease Inhibitor Cocktail (cOmplete) stock solution was added to 80ml of HEGX Buffer. cOmplete was added directly before use and stored at 4°C.
10% PEI (12.5ml)	2.7g of 50% PEI was mixed with 5 ml H_2O and 1.5 mL 37% HCl (on vortex); then 2 mL 5 M NaCl, 0.25 mL 1 M HEPES, pH 7.2, 7.5 µL 0.5 M EDTA and 0.25 mL 10% Triton X-100 were added to the PEI mixture and adjusted to 12.5 ml with H_2O in the cylinder. pH was checked with pH-Fix strips (Fisherbrand) and the solution stored at RT (pH was in the range of 7-8).
2x Tn5 Exchange Buffer (50ml)	5ml 1M HEPES-KOH (pH 7.2), 2ml 5M NaCl, 20µl 0.5M EDTA, 1mL 10% Triton X-100, 10ml 100% Glycerol and 100µl 1M DTT were mixed and adjusted with H_2O up to 50ml. The solution was stored at 4°C.
1x TAE (1l)	20ml 50x TAE were mixed with 980ml H_2O and stored at RT.

Name	Recipe
10x STE (10ml)	2ml 5M NaCl, 1ml 1M Tris-HCl (pH 8.0), 200µl 0.5M EDTA, were mixed and adjusted with H_2O up to 10mL. The solution was stored at 4°C.
2x Oligo Loading Buffer (10ml)	5mg of Orange G was mixed with 0.4ml of 0.5M EDTA and 9.5ml of 100% Formamide. The solution was stored at RT.

Methods

Tn5 Production

The pTXB1-Tn5 construction along with technical documentation was kindly provided by the group of Rickard Sandberg. The plasmid was delivered from Addgene's repository as transformed bacteria in stab culture format. The expression of the protein was performed as previously described [Picelli et al., 2014]. The modified purification protocol was conducted according to the literature by using chitin magnetic beads instead of chitin magnetic column.

Bacterial Stab

A LB petri dish with 100µg/ml Ampicillin was streaked from the bacterial stab and incubated overnight at 37°C. Two single colonies were picked and grown in 2 x 5mL of LB medium (Ampicillin 100µg/ml) in the 37°C thermoshaker overnight. The liquid bacterial culture was then used for purification of the plasmid and glycerol stock solutions.

Glycerol Stocks

To store the characterized, transformed bacteria, 0.5ml of a liquid overnight culture was mixed with 0.5ml 50% glycerol in a 2ml cryotube and stored at -80°C. From these frozen glycerol stocks, fresh overnight cultures were inoculated.

Plasmid DNA Preparation

Overnight bacterial culture 4.5ml x 3 per clone was set up from single bacterial colonies in LB-medium supplemented with ampicillin. The plasmid DNA was prepared using QIAprep Spin Miniprep Kit and eluted in 50µl of Elution Buffer. The concentration of the plasmid was evaluated with Qubit-iT dsDNA HS Assay Kit.

Sequencing

The insertion of Tn5 in two clones was verified by Sanger sequencing from both sides (T7 promotor/Seq_Rev_Tn5). After comparison with Tn5 sequence kindly provided by Rickard Sandberg's laboratory the plasmids were used for protein expression (Supplementary Table 2).

Protein expression

Along the experimental procedure, cell culture, supernatant, pellet or chitin magnetic beads aliquotes (40µl) were taken, frozen and kept in 1.5ml SafeLock tubes. After collecting, all samples were analysed on 10% SDS-PAGE (Coomassie staining).

Transformation T7 Express lysY/Iq Competent E. coli cells

Competent *E. coli* cells were defrosted on ice and plasmid DNA (about 100ng) was added. After mixing the contents gently, the suspension was stored on ice for 3 min. Bacteria were heat shocked for 10 sec at 42°C and put back onto ice for 5 min. 950µl of room temperature SOC medium was added and the culture was incubated for 1 hour at 37°C with vigorous shaking (200rpm). The selection LB plates were prewarmed at 37°C. After incubation time 50µl of the culture was used to prepare 4-fold and 6-fold dilutions in SOC medium. 50µl of cell dilutions was spread onto an LB-Amp plate directly. Plates were left at RT until the liquid was absorbed and incubated overnight at 37°C.

Overexpression

Two individual C3013 colonies were picked up and the cells were grown in 50ml of liquid LB medium with Ampicillin at 37°C overnight. 10 ml of overnight culture was used to inoculate 1l of liquid LB with antibiotic. The cells were grown for about 6 hours with shaking (250rpm) at 25°C to optical density OD 0.664. At OD about 0.664 the culture was induced with 250µl of 1M IPTG and grown for an additional 4 hours. The dynamics of bacterial growth for two clones is shown (Table 1):

Table 1 Optical density of cell biomass

Time	OD Clone-1	OD Clone-1
14:15	0.072	0.073
15:50	0.259	0.251
17:40	0.684	0.664
21:30	2.116[I]	2.277[I]

The value with I corresponds to OD of induced cell culture.

After protein induction with IPTG the cell culture was centrifuged for 15 minutes at 5000g at 4°C in 250 ml autoclaved centrifugation bottles. The culture was separated into portions and four cycles of centrifugation were performed according to the volume limit of centrifugation bottles. The pellets were resuspended on ice in 40ml of cold TEX buffer and the final suspension transferred to 50 ml centrifuge tubes (2 x 40mL per clone). To simplify the process the pellets were first vortexed for about 10 minutes and then the cell debris was dissolved in cold HEX Buffer by pipetting up and down. The resuspended cell culture was centrifuged again for 15 minutes at 5000g at 4°C in 50ml tubes. The overall pellets were frozen in two tubes and kept at -80°C overnight.

Tn5 purification

Preparation of Chitin Magnetic Beads

Chitin magnetic beads were prepared on ice according to the manufacturer's instructions. 10ml of the beads were washed in portions with a 10-fold volume of cold HEGX buffer until final dilution in original volume.

Binding

The pellets were thawed on ice for 20 min, vortexed for 10 min and resuspended in 40ml of cold HEGX cOmplete buffer by pipetting and avoiding foaming. The resuspended cells were sonicated by the use of Microfluidizer Processor (~80psi, 4x10 presses, circulating). The cell debris was removed by centrifugation at 14000 x g for 30 min at 4°C. *E. Coli* genomic DNA was precipitated by adding 1.2ml of 10% PEI dropwise while stirring to a 40ml of the

supernatant followed by centrifugation at 15000 x g for 10 min at 4°C. 80ml of the lysate was mixed on ice with 10ml previously prepared chitin magnetic beads (total volume 45ml per tube). Binding was performed for 4.5 hours at 4°C under rotation. After incubation, the beads were collected on ice by use of a 15ml tube magnetic rack and washed by 5 iterations with the overall volume of 280-300ml of cold HEGX buffer (5 min incubation time per each beads separation).

Cleavage

The cleavage of Tn5-transposase from the intein domain was performed by resuspending the washed beads in 20ml of cold HEGX buffer with 50mM DTT followed by the rotation for 36h at 4°C.

Elution

After the cleavage the suspension of beads was slightly mixed and separated on the 15ml tube magnetic rack. A supernatant was collected and transferred to a new 50ml tube. To remove the residual amount of beads the supernatant was exposed to the second separation on a 2ml tube magnetic rack. The overall supernatant was collected on ice in a new 50ml tube. Analogously, the magnetic beads were eluted in 8ml of cold HEGX buffer 5 additional times. Tn5 concentration was measured in each fraction using the Qubit Protein Assay Kit according to the manufacturer's instructions (Table 2).

Table 2 Concentration of Tn5 Transposase in elution fractions

Fraction	Concentration (µg/µl)	Volume (ml)
1	0.452	18
2	0.442	5
3	0.426	5
4	0.422	5
5	0.426	5
6	0.442	5
1-6	0.444	43

10µl of fraction volume was used per measurement

Protein concentration and buffer exchange

After the measurement of the Tn5 concentration, all protein fractions were combined and exposed to the final beads separation on a 2ml tube magnetic rack; residual amount of beads in the supernatant may clog concentrators. The protein concentration was carried out using Amicon Ultra-4 Centrifugal Filter Devices (30K). The total fraction volume (43ml) was concentrated to 200µl at 7500g (fixed-angle rotor, 2 filter units) for 45 min at 4°C. However, excessive concentration leads to some surface denaturation, which lowers the Tn5 specific activity. 200µl of the protein solution was dialyzed versus two changes of 6600µl of 2X Tn5 Exchange Buffer to 125µl x 4 of final protein solution at 3200g for 1.5h at 4°C (swinging-bucket rotor; 4 filter units). After dialysis, the Tn5 solution was mixed at a ratio 1:1 with 80% glycerol. The final concertation of Tn5 Transposase was measured as described above (1:10 dilution); then the protein stock solution was aliquoted and stored at -80°C (or -20°C for working aliquots).

Transposase activity assay

To check tagmentation activity of in-house Tn5 transposomes only Illumina type A adapter was used. The oligonucleotides Tn5ME-A and Tn5MErev were annealed and the 5µM

glycerol solution was prepared as described [Wang et al., 2013]. A 1:10 dilution of home-made Tn5 stock solution was prepared in 1X Tn5 Exchange Buffer, 50% glycerol and 7.5µM working solution was obtained. The Tn5 transposomes were assembled by mixing 10µl of 5µM adapter A and 10µl of 7.5µM Tn5 Transposase working solution and then incubating the mixture for 30 min at 25°C. Analogously, transposome assembly was prepared with a home-made Tn5 transposase aliquot from our colleagues (Dr. Alisa Fuchs) to use it as a control sample. Tn5 transposome mixtures were kept on ice, when used in the same day or stored at 20°C. pKSII plasmid linearized with EcoRI was applied as a tagmentation substrate. An exonuclease digestion was performed in 20µl of reaction volume by incubation 1µg of plasmid with 15 units of EcoRI in 1X Buffer H at 37°C for 2 h. The enzyme was heat inactivated at 65°C for 20 min and the reaction purified with QIAquick PCR Purification Kit; concentration of cut plasmid was assessed using Qubit dsDNA BR Assay Kit according to the manufacturer's datasheet. Tn5 transposome assembly was titrated by volume to generate tagmentation of 50ng plasmid DNA in 1X TB Buffer in total volume of 30µl for 1h at 37°C. A 1.2µl of 2% SDS solution was added to each reaction including controls, which was then incubated at 55°C for 7 min to inactivate the Tn5 and release it from the DNA. The samples were then analyzed by 1.2% gel electrophoresis.

Fragmentation Efficiency Assay (FEA)

Transposon

The Tn5 transposon end adapters used for the FEA assay are the Illumina NGS libraries preparation scheme adapters. 80µM Tn5ME-A and 80µM Tn5ME-B were annealed to 80µM Tn5MErev at a ratio of 1:1, forming 40µM Tn5ME-A/rev and Tn5ME-B/rev oligonucleotide duplexes. The oligonucleotides were annealed overnight in total volume of 40µl by natural temperature decrement, incubating the tube in the lab glass filled with 100°C water. Adapter duplexes were then mixed and diluted two times with 100% glycerol to obtain 20µM Tn5ME-A/B adapters (each adapter 10 µM), 50% glycerol solution. The Tn5ME-A/B adapter glycerol stock was stored at -20°C.

Transposome assembly

For transposome assembly, a 20µM Tn5 solution was prepared in 1X Tn5 Exchange buffer, 50% glycerol. The same aliquot of Tn5 solution was used for all FEA transposome assemblies. Tn5 transposome were prepared by mixing equal volumes of 20µM Tn5 enzyme and 20µM Tn5ME-A/B adapters incubating the mixture for 1h at 25°C. The resulting

assembly was stored at -20°C. For evaluation of the in-house produced Tn5 transposomes, the TDE1 Tagment DNA enzyme from Illumina Nextera Rapid Exome Kit was used.
In-house transposome assemblies were added to tagmentation reactions without removal of residual adapters and free Tn5 molecules. It was, therefore, not possible to determine the actual concentration of transposomes. In the Illumina TDE1 mixture, the concentration of transposomes is also not indicated. Therefore, for both the in house prepared and commercial mixtures, the volume was used as a measure of the amount of the transposome assembly used in tagmentation reactions.

Tagmentation template

A commercially available pUC19 plasmid (2686bp) was used as a tagmetation template. The plasmid was linearized with EcoRI restriction endonuclease as follows: 1 µg of plasmid DNA was incubated with 15 units of EcoRI enzyme in 20µl of 1X Buffer H at 37°C for 2h. The endonuclease was heat inactivated at 65°C for 20 min and the reaction purified with QIAquick Gel Extraction Kit. Digestion efficiency was checked on a 1% agarose gel and DNA concentration measured on a Qubit Fluorimeter using the Qubit dsDNA BR Assay Kit pUC19/EcoRI, 22ng/µL was used in all tagmentation reactions. The same plasmid DNA diluted in EB Buffer down to 0.5ng/µl was used as a spike-in. For spike-in experiments a commercially available human genomic DNA (200ng/µl) was used. The DNA concentration was estimated using Qubit DNA BR Kit according to manufacturer's protocol. Three independent measurements were performed and a mean value (145ng/µl) was used. The integrity of genomic DNA was proved by a single band on 1% agarose gel. The genomic DNA solution was aliquoted by 50µl and stored at -20°C.

EB Buffer: 10mM Tris-HCl pH 8.5

Tagmentation

Tagmentation reactions were performed in 1X TB Buffer with 50ng of pUC19/EcoRI in 25µl. For the spike-in experiment, 0.5ng of the plasmid was added to 50ng of human genomic DNA for tagmentation in total reaction volume of 50µl. For experiments involving transposome assembly dilutions, those dilutions were first prepared separately from the stock assembly in 0.5X Storage Buffer and 50% Glycerol and then equal volumes of transposomes were added to the tagmentation reactions run in parallel. Negative tagmentation control without transposase was always performed in parallel. Tagmentation temperature varied depending on the experiment. For evaluation of the fragmentation efficiency assay reproducibility and

tagmentation bias, tagmentation was performed for 1 h at 37°C. Longer time was selected so that small time differences related to tubes handling would have no influence on the results. For the comparison of the in-house Tn5 tranposomes and Illumina TDE1 mixture and for tagmentation of human genomic DNA with a spike-in, tagmentation reactions were performed for 10 min at 58°C, according to the conditions recommended by the Illumina Nextera protocols for the NGS library preparation. Tagmentation reactions were stopped by adding 2% SDS solution (to the final concentration 0.08%) and incubating for 7 min at 55°C. Tagmentation reactions were purified with AMPure XP Beads which provides efficient removal of the DNA molecules of smaller sizes (cut-off is adjustable). Purification was conducted as the 10µM adapters in the in-house transposome assembly inhibit PCR reactions, especially when more than 30pmol are taken per 25µl of tagmentation reaction. Purification was carried out according to the manufacturer's protocol, with the following settings: beads were added to the Tn5 reactions at a 0.8:1 ratio; beads were washed with 70% ethanol; DNA was eluted in EB Buffer, in the same volume as taken for purification. In spike-in experiment beads were eluted in 25µl of EB Buffer.

qPCR

Aliquots of the purified tagmentation reactions were analysed by real-time PCR: 1/150 was used for evaluation of the plasmid-only tagmentation; 1/300 for Illumina/in-house comparison tests; and 1/25 - for the plasmid spike-in experiments. qPCR was conducted in StepOne Real-Time PCR machine using the SYBR Green PCR Core Reagents and 1 unit of Immolase per reaction. Following temperature profile conditions were applied: 95°C for 10 min followed by 40 cycles of 95°C for 15 sec, 63°C for 15 sec and 72°C for 60 sec (PCR products < 1 kb), for 70 sec (for 1240bp PCR product) or 120 sec (for 2248bp PCR product). Each reaction contained 0.5µM forward and 0.5µM reverse primers, in a final reaction volume of 20µl.

Visual fragments size analysis

Fragments sizes were checked loading half of the purified tagmenation reaction volume on 1% UltraPure Agarose gel. Electrophoresis was performed in 1X TAE Buffer at 120V for 2.5h. Alternatively, 1µl of the purified tagmenation reaction was checked on Agilent 2100 Bioanalyzer using High Sensitivity DNA Kit.

PITS Library Preparation

Transposome assembly

The Tn5 transposon end type A adapter oligonucleotides Tn5ME-A (top) and p_026 (bottom) were annealed by mixing 10µl of each oligonucleotide at the concentration of 80µM. Then 60µl of RNase-free water was added to the oligonucleotide duplex to obtain 80µl of 10µM solution. The thermal profile of Tn5 adapter annealing is shown (Table 3):

Table 3 Thermocycling conditions for adapter annealing

Denature	Anneal	Ramp to 26°C	Hold
95°C, 3 min	70°C, 3 min	0.1°C/s	26°C, infinite

100µl of 100% glycerol was heated to 90°C followed by transferring a 80µl glycerol aliquot into 0.5ml tube (hot glycerol can be exactly pipetted) and then cooling it down to RT on ice for 3 min. 80µl of oligonucleotide duplex was added to the equal volume of pre-cooled glycerol. 20µM Tn5 Transposase working solution was prepared on ice by dilution of a Tn5 stock in 1X Tn5 Exchange Buffer, 50% glycerol. The Tn5 transposome was assembled by mixing 80µl of 20µM in-house Transposase and 80µl of 5µM Tn5 adapter in glycerol (4:1 by quantity) followed by incubation at 25°C for 21h. The transposome assemblies were stored at -20°C for at least 1 month.

Phage Lambda genomic DNA

A commercially available DNA of phage phage (500ng/µl) was diluted in RNase-free water to a working concentration of 36ng/µl. The concentration of DNA dilution was estimated using Qubit DNA HS Kit according to manufacturer's protocol. Three independent measurements were performed and a mean value was used. The integrity of genomic DNA was proved by a single band on 1% agarose gel. The working DNA solution was aliquoted by 50µl and stored at -20°C.

Tagmentation

The tagmentation was carried out using the reaction premix for 4 samples (Figure 14).

50ng of Lambda genomic DNA (1 and 2) was tagmented by mixing 15µl of Tn5 Transposomes, 10µl of 5X TB Buffer and water to a total volume of 50µl and then by

incubating the mixture at 55°C for 8 min. 50µl of 40mM EDTA was added to stop the reaction. The DNA-free (3) and Tsome-free (4) negative control samples were processed in the same manner except for replacing Tsomes or phage DNA with RNase-free water.

2µl of 2% SDA was added to 50µl of size control tagmentation reaction (1) followed by incubation the mixture at 55°C for 7 min. The reaction product was purified on AMPure XP DNA beads at ratio of 1:2.5 and eluted in 25µl of RNase-free water. 20µl of reaction volume was loaded on 1.1% agarose and run at 100V for 2h; 1µl of purified sample was used for Bioanalyzer to monitor the size distribution of tagmented DNA by 15µl of Tsomes.

Dilution

Commercially available tRNA was used to prepare 2ng/µl working solution. The manufacturer's supplied concentration was confirmed using the Nanodrop 1000 Spectrophotometer (Thermo Scientific). The working solution was aliquoted and stored at -20°C. 5µl of tagmentation reactions were diluted with 2ng/µl tRNA to obtain 150µl of 10^3molecule/µl solution. The overall volume was aliquoted in 0.2ml 8-strips by 4.5µl and stored at -20°C.

Oligo replacement and gap-filling reaction

4.5µl aliquots of tagmented lambda DNA and appropriate controls were thawed on ice and mixed with 0.5µl of 0.1% SDS by gentle pipetting to remove Tn5 protein from the DNA. For oligo replacement and annealing, 1µl of 1µM replacement oligonucleotide p_028, 1µl of 10X Ampligase Buffer (enzyme supplement), 1µl of 2.5mM dNTP's mix (each) and 1µl RNase-free water were added to 5µl of released tagmented DNA and mixed by pipetting. 9µl of reaction was incubated in the thermocycler as follows (Table 4):

Table 4 Thermocycling conditions for oligo replacement and annealing

Denature	**Anneal**	**Ramp to 37°C**	**Hold**
50°C, 1 min	45°C, 10 min	0.1°C/s	37°C, infinite

The for a gap-repair reaction a combined action of T4 DNA Polymerase and Ampligase was used as previously described [Wang et al., 2013]. 0.5µl of 100u/µl Ampligase and 0.5µl of 3u/µl of T4 DNA Polymerase were mixed on ice in a separate 0.2ml single tube (the volume was scaled according to the amount of reactions). 1µl of gap-repair mix was added to 9µl of

oligo replacement mixture, while the tube remained in the thermocycler; mixed gently by pipetting and incubated for 30 min at 37°C followed by cooling to 4°C. No purification was carried out after this step.

The activity of Ampligase and T4 DNA Polymerase individually or both enzymes in combination was evaluated by their titration using oligonucleotide compositions as model substrates. A 1:2 serial dilution was performed as described [Adams, 2003]. Two ligases, T4 DNA Ligase (400 units) and Ampligase (100 units), per 10µl reaction were compared under T4 DNA Polymerase 1:2 serial dilutions, starting with 1.5 units of polymerase. Model duplexes were prepared by annealing three oligonucleotides (p_015, p_028 and p_029) to generate a 9nt gap.

The final SDS concentration in the gap repair reaction was estimated by titration of 0.2% SDS and controlling the yield of a ligated product under established working conditions. The titration was scaled for 10 reactions. 10µl of each nucleotides p_015, p_028 and p_029 in 20µM were mixed and heated to 95°C for 1 min, followed by a ramp to 4°C with 0.1°C/sec. 30µl of annealed oligonucleotides, 10µl of 10X Ampligase Buffer, 10µl of 10mM dNTP's mix (each) and 30µl of RNase-free water were mixed on ice. The total reaction mixture was dispensed by 10µl into 0.2ml 8-strip with a double volume in the first tube. 0.5µl of 0.2% SDA solution was added to the 16µl in the first strip tube and 1:2 serial dilutions were performed. 10µl of 100u/µl Ampligase and 10µL of T4 DNA Polymerase were combined on ice. 2µl of enzyme mix was added to 5 dilution points. Three control samples were planned: Ampligase only, T4 DNA Polymerase only or absence of both enzymes. The samples were mixed gently and incubated in the thermocycler at 37°C for 30 min. 10µl of PAAG Buffer was added to each strip tube, mixed, heated to 95°C for 5 min and then immediately transferred to an ice bath. 10µl reaction aliquots were examined by 10% denaturing polyacrylamide gel electrophoresis. The gel was stained for 30 minutes with a 1:10000 dilution of SYBR Green II in 1X TBE.

Ampligase 10X Reaction Buffer: 200 mM Tris-HCl (pH 8.3), 250 mM KCl, 100 mM $MgCl_2$, 5 mM NAD, and 0.1% Triton® X-100.

Amplification

Real-time PCR reagents with adapter-specific primers that are compatible with the sequencing oligonucleotides (TruSeq dual-index sequencing, paired-end; Illumina) were used to amplify the library. 10µl of gap-repaired product, 2.5µl of 20µM primer p_032, 2.5µl of 20µM primer p_033, 20µl of 25mM $MgCl_2$, 20µl of dNTP's mix (with dUTP's), 20µl of 10X SybrGreen,

1µl of 5u/µl Immolase DNA Polymerase and 124µl of RNase-free water were combined on ice in 0.2ml PCR single tube. The final reaction volume was 200µl. The reaction was heated to 95°C for 10 min and then thermocycled 24 times at 95°C for 15 sec, 62°C for 15 sec, and 72°C for 30 sec. The amplified library was cooled to 4°C and cleaned on MinElute columns using QIAquick PCR amplification Kit according to manufacturer's instructions. The column was eluted with 10µl of RNase-free water. The volume of the purified library was adjusted to 5µl by gradual evaporation at 37°C in the thermocycler.

Flow cell loading and sequencing

The loading concertation of DNA library was evaluated by qPCR analysis. The MiSeq flow cell was loaded as described in MiSeq System Denature and Dilute Libraries Guide (www.illumina.com) except for several modifications. 5µl of 1.2nM library was denatured by 1µl of 0.6N NaOH. 6µl of the denatured library was mixed with 594µl of hybridization buffer (HT1). The PITS library was sequenced using PE 33 Index Kit MiSeq v1.

Sequencing analysis

FASTQ files were generated with Illumina BCL2FASTQ Conversion Software. Before alignment, preliminary quality control of FASTQ files was performed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) software. Lambda phage reference genome sequence in FASTA format was taken from NCBI website: https://www.ncbi.nlm.nih.gov/nuccore/215104?report=fasta. Alignment was performed using Bowtie 2 software (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). Index for the Lambda phage reference genome was generated using FASTA file from NCBI. Alignment of paired-end reads was performed using default settings of the aligner resulting in a SAM file with aligned reads. Obtained SAM file was further converted into BAM file, sorted and indexed using Samtools software (http://www.htslib.org/). Duplicates were detected with Picard tools (https://broadinstitute.github.io/picard/). Non-duplicated read pairs (fragments) that are properly (both reads of the fragment are aligned to the same strand with a total length of the fragment no more than 500bp) and uniquely aligned were then identified using Samtools. Using homemade python script, fragments identified in the previous step were utilized to reconstruct scaffolds, representing sequences of directly adjacent fragments. Two fragments were considered as adjacent, if 9 bases at the end of one fragment were identical to 9 starting bases of another fragment.

Preparation of Alternative Transposon structures

Synthetic Lampion Transposon

Three oligonucleotides p_008, p_009 and p_010 were used for preparation of a Tn5 lampion-like adapter structure by ligation of 2 duplexes. The oligonucleotide p_010 was first phosphorylated at 5´ end. 1000pmol of p_010 was mixed with 2µl of 1X T4 PNK Buffer, 2µL of 10mM rATP, 0.5µl of 10u/µl T4 PNK and RNase-free water to total volume of 20µl and then by incubating the mixture in the thermocycler for 1h at 37°C. The T4 PNK was deactivated by heating the mixture for 20 min at 65°C. The p_010 phosphorylated oligonucleotide was used in the ligation reaction without additional purification and stored at -20°C. Prior to ligation, two types of duplexes were prepared. The first duplex was prepared by mixing 50pmol of p_008 (biotinylated), 55pmol of phosphorylated p_010, 5µl of 10X T4 Ligase Buffer, 3.3µl of 60% PEG and RNase-free water. The oligonucleotides were annealed in total volume of 50µl by natural temperature decrement, incubating the tube in the lab glass filled with 100°C water. The second duplex was prepared in the same manner except by adding 50pmol of p_009 instead of oligonucleotide p_008. 30µl of each duplex were combined and 6µl of 400u/µl T4 DNA Ligase was added to the mixture followed by incubation for 2h at 25°C. The efficiency of ligation reaction was assessed by loading a reaction aliquot (5pmol) on 2% agarose (GelRed staining) and 6% polyacrylamide denaturing gel (SYBR Green II staining). The lampion-like adapter was stored either at 4°C or at -20°C. The adapter was purified using QIAquick Gel Extraction Kit, AMPure Beads according to manufacturers' instructions or by TAE/PEG 15% gel extraction technique as described [http://molbiol.ru/protocol/08_01.html].

Preparation of PCR Products of Different Lengths

Three different PCR products were prepared by two-round amplification. Primers were designed to amplify the following fragment lengths: 158bp (p_005/p_006), 258bp (p_011/p_012) and 495bp (p_013/p_014). All oligonucleotides were phosphorylated at 5' end and used without additional purification in PCR. 1000pmol of each primer was incubated in the thermocycler with 5 units of T4 PNK in 1X T4 PNK Buffer with 1mM rATP in 20µl reaction for 1h at 37°C. The T4 PNK was deactivated by heating the mixture for 20 min at 65°C. For the first round of PCR mouse genomic DNA was used as a PCR template. Different lengths of double-stranded DNA were amplified from 100ng of DNA in 1X PCR Buffer with 0.2mM dNTP's mix, 2.5mM $MgCl_2$ and 0.5µM of forward and reverse primers by adding 25

units of AmpliTaq Gold DNA Polymerase. The total volume of each reaction was 50µl. The reaction mixtures were heated to 90°C for 10 min and then thermocycled 31 times at 90°C for 15 sec, 65°C for 15 sec and 72°C for 15 sec, followed by a final extension of 30 sec at 72°C. PCR products were purified using QIAquick Gel Extraction Kit and their concentrations were measured on a Qubit Fluorometer using the Qubit dsDNA BR Assay Kit. 0.2ng of purified PCR products were used as templates for the second round of PCR. 32 reactions for each fragment length were performed as described above except for 2 changes: 1.25 units of Immolase instead of AmpliTaq Gold DNA Polymerase were used per 50µl reaction; each reaction was thermocycled for 29 times. The resulting PCR products were again gel purified and end-repaired using 40µl of End Repair Master Mix per 100µl of final volume (1100ng per reaction) and incubating the mixture for 1h at 30°C. Blunt-ended PCR products were cleaned-up by using AMPure beads in 1:1.8 ratio. The concentrations of PCR products were measured using the Qubit dsDNA HS Assay Kit.

10X PCR Buffer: 100mM Tris-HCl pH 8.0, 500mM KCl, 0.1mM EDTA, 0.5% Tween20.

Transposon Insertion with Epicentre Transposase

Two tests were performed to evaluate the activity of commercially available Tn5 Transposase from Epicentre. The 158bp double-stranded PCR product was used as Tn5 transposon along with EZ-Tn5 KAN-2 Transposon. For the first test two ratios of target DNA to transposon were used: 1:5 and 1:15. Transposition was performed in 1X EZ-Tn5 Buffer by mixing 0.1pmol of pKSII/EcoRI plasmid DNA (2961bp), 0.5 or 1.5pmol of 158bp Transposon and 0.5 units of EZ-Tn5 Transposase and incubating the reaction for 2h at 37°C. Linearized pKSII plasmid DNA was prepared as described above (Transposase Activity Assay Section). The control reaction was performed using Epicentre target DNA (pUC19/3.4) and 1221bp EZ-Tn5 KAN-2 Transposon in ration 1:5. The volume of each reaction was 10µl. The influence of transposition conditions on the target DNA were checked by Tson/Tsase-free reactions. Transposition was stopped by adding 1µl of EZ-Tn5 Stop Solution (1% SDS) and heating the reaction mixture for 15 min at 70°C. The samples were then analyzed by 1% agarose gel electrophoresis. The second test was performed under the same conditions except for using end-repaired (blunt-ended) 158bp Transposon and only 1:5 target DNA to Transposon ratio. 158bp Transposon and EZ-Tn5 KAN-2 Transposon insertion reactions were carried out using pKSII/EcoRI target DNA and pUC19/3.4 target DNA correspondingly.

pUC19/3.4 Control Target DNA (0.1µg/µl): 3.4-kb HpaII fragment of bacteriophage DNA cloned into the AccI site of pUC19, in TE Buffer (10 mM Tris-HCl pH 7.5 and 1 mM EDTA).

EZ-Tn5 10X Buffer: 0.50M Tris-OAc pH 7.5, 1.5M KOAc, 100mM $Mg(OAc)_2$, 40mM spermidine.

Transposon Insertion with In-House Tn5 Transposase

Insertions of three full transposons with different lengths (158bp, 258bp and 495bp) were tested with In-house Tn5 Transposase. Linearized plasmid pKSII was used as a target DNA. Transposition reactions were performed in 1X EZ-Tn5 Buffer by mixing 0.1pmol of pKSII DNA (2961bp) with 10-fold excess (1pmol) of each transposon and adding 4pmol of In-house Tn5 Transposase (4µg/µl working solution) and incubating the reaction in the thermoblock for 2h at 37°C. Linearized pKSII plasmid DNA and Transposons were prepared as described above (Transposase Activity Assay Section; Preparation of PCR Products of Different Lengths Section). Transposition was stopped by adding 1µl of EZ-Tn5 Stop Solution (1% SDS) and heating the reaction mixture for 15 min at 70°C. The samples were then analyzed by 1.1% agarose gel electrophoresis.

EZ-Tn5 10X Buffer: 0.50M Tris-OAc pH 7.5, 1.5M KOAc, 100mM $Mg(OAc)_2$, 40mM spermidine.

RESULTS AND DISCUSSION

Preface

The aim of the PhD project was to determine principal issues of experimental realization of the paired indexing of tagmentation sites (PITS) approach, develop a relevant PITS protocol and evaluate feasibility of the approach.

Tagmentation is a starting point of PITS. The method's current performance and its future prospects strongly depend on the properties of Tn5 transposase. For both establishing an initial protocol and testing alternative methodological solutions generous amounts of Tn5 transposase are required. Commercial enzymes are available either as part of EZ-Tn5 insertion kits (Epicentre), which are very expensive, or as complexes with oligonucleotides (Illumina), which restricts their use to the manufacturer's protocols. Therefore, along with the main stream of the project, the in-house Tn5 transposase was prepared.

Tn5 unit definition is not practical for NGS applications where not the Tn5 itself, but rather transposomes – complexes of an enzyme with transposons – act as activity units. To be able to adjust activity of Tn5 transposome batches relative to each other and also to compare transposome performance in different experimental settings, we developed an assay to characterize tagmentation activity of transposomes.

Thus, the practical work performed during PhD clearly splits into three blocks: development of the tagmentation sites coding approach, preparation of the in-house Tn5 enzyme, and development of Tn5 transposomes DNA fragmentation efficiency assay.

Tagmentation sites indexing approach for contiguity-preserving sequencing

The general principle of the PITS approach is shown in Figure 10A. The main challenge in the development of the experimental protocol is to prevent losses of tagmentation products and, ideally, get all of them sequenced. A missing fragment means a guaranteed gap in the subsequent DNA assembly. The more gaps the shorter fragment scaffolds would be. This issue requires efficient generation of tagmentation products in a size range appropriate for a sequencing platform used and sets constraints on PITS sequencing library preparation. Tn5 tagmentation is known to produce a wide size range of fragments and is not random [Adey et al., 2010; Rykalina et al., 2017]. Some fragments would definitely be out of sequencable size. However, tagmentation is a standalone step and has a promising adaptability potential.

Currently, researchers are working to obtain a less sequence-dependent transposase and to optimize tagmentation conditions to decrease length variability of the resulting fragments [US 2015/0291942 A1; Kia et al., 2017]. Below we also discuss the options to improve the yield of sequencable tagmentation products with the available Tn5 enzyme. In our view, library preparation in particular, as well as its efficient loading on a sequencer, presents a fundamental challenge for the performance of the PITS approach. Therefore, experimental work on PITS was organized to address these issues and included: (i) development of a dedicated protocol for NGS library preparation; (ii) selection of a relevant test system; (iii) proof-of-principle experiment; (iv) evaluation of primary results, revealing critical points and bottle-necks and (v) identification of further development strategies.

NGS library preparation from amol amounts of DNA and non-residual loading on Illumina sequencing flowcell

Suppose, we run a PITS sequencing library on a single lane of Illumina HiSeq2500old sequencer and aim to sequence all initial tagmentation fragments. DNA sequence of which length could we theoretically reconstruct? The following estimate illustrates the challenge of a sequencing library preparation for the PITS approach.

The expected average cluster number per HiSeq2500old lane is 150mln (15×10^7), so 15×10^7 library molecules get sequenced. Not all molecules of the loaded library hybridize to the flowcell surface, so, since we need to retrieve the whole variety of tagmentation fragments, which were converted to library molecules, each fragment should be present in several copies to increase the chance of being sequenced. Let's assume 10 copies of each fragment are sequenced – then a library solution loaded on a flowcell should contain copies of 15×10^6 initial tagmentation fragments. Since we want to be able to link all fragments using paired indices, we should start with 15×10^6 tagmentation fragments and preserve them through library preparation. Counting 300bp as an average tagmentation fragment size, the total length of DNA taken for tagmentation should be 4.5Gb which corresponds to ~5pg of DNA. Table 5 shows the results of analogous calculations for several other types of Illumina sequencers. Summarizing, to execute PITS sequencing library preparation protocol, amol amounts of short fragments should be converted into sequencing library molecules with minimal losses and sequenced.

Table 5 Estimation of the PITS protocol starting DNA amounts basing on the final sequencing volume

Desired Sequencing Volume	Amount of Expected Clusters	Number of Library Molecules to Be Sequenced (~10x coverage)	Estimated Length of the PITS starting DNA*	Estimated weight of the PITS Starting DNA
1x MiSeq flowcell reagent kit v.2	15×10^6	$1{,}5 \times 10^6$	450Mb	~0.5pg
1x HiSeq2500old lane reagent kit v.3	150×10^6	15×10^6	4, 5Gb	~5pg
1x HiSeq2500 lane reagent kit v.3	220×10^6	22×10^6	6.6Gb	~7pg
1x HiSeq4000 lane	285×10^6	$28{,}5 \times 10^6$	8,6Gb	~9.5pg

*Calculated for 300bp average size of the tagmentation fragments

We developed a protocol allowing to efficiently prepare sequencing libraries from amol amounts of tagmentation fragments – post-tagmentation ultra low input (PTULI) sequencing libraries protocol. The laboratory workflow and the schematic representation of the developed PTULI procedure are shown in Figure 11. Briefly, an aliquot of tagmentation reaction containing the required amount of DNA fragments is taken for library preparation. Transposase removal, exchange of an unligated transposon strand with a tailed adapter oligonucleotide, filling in 9nt gaps, left by transposase and amplification, are sequentially performed in a single tube, without intermediate aliquoting or purification steps. The clean-up procedure is carried out after amplification. Since the sequencing platform available for this project was Illumina MiSeq, the protocol uses Illumina-specific technical sequences.

Below we describe in detail each step, as well as the process of loading PTULI sequencing library on a flowcell.

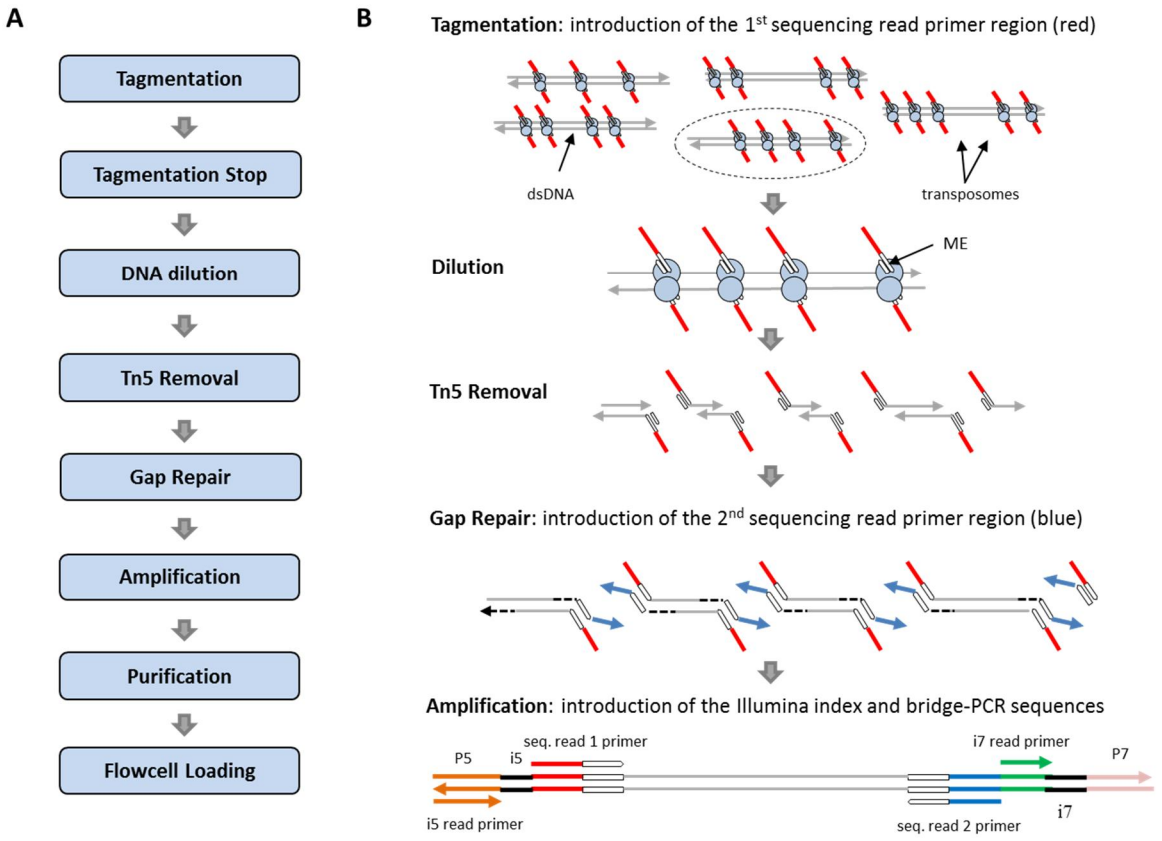


Figure 11 PTULI library preparation protocol suitable for PITS approach. (A) Laboratory procedure workflow. (B) Scheme of the method. Transposomes are depicted as double circles, each circle with partly double arrows, corresponding to transposase dimers bound to transposon adapters. Transposase recognition sites (ME) are shown as empty double arrows. Colours of Illumina sequences are explained on the scheme. Dashed lines mark the sequences synthesized during the gap-filling reaction.

Tagmentation

Tagmentation reaction is performed on 50ng of DNA, far larger amount than deemed reasonable for PITS as calculated above. Recently it has been shown that during tagmentation, after inserting breaks into dsDNA and ligating adapters to the arising 5´ ends, Tn5 transposase dimers do not dissociate from the DNA molecules and hold the fragments together [Amini et al., 2014]. Thus, during tagmentation contiguity of the original long molecules is preserved, which allows to sample the required amounts of DNA from already tagmented material. This possibility makes tagmentation step as controllable as in standard tagmentation based NGS protocols (e.g. in Illumina Nextera kit procedures). For 50ng of DNA, in contrast to 5pg, the resulting fragment sizes may be analysed on a gel or Bioanalyzer, so tagmentation conditions for PITS can be initially tuned (Supplementary Figure 3). Besides, ng DNA amounts can be measured more precisely than pg amounts, which is crucial for reproducibility of tagmentation, known to be sensitive to the ratio of DNA and transposomes [www.illumina.com].

Transposomes are assembled with transposon end type A adapters, which are prepared from two oligonucleotides: Tn5ME-A (top) and p_026 (bottom). The longer (top) Tn5ME-A oligonucleotide contains Illumina technical sequence corresponding to the 1st sequencing read primer (marked red in Figure 11). Top oligonucleotides are ligated to the 5´ ends of the arising tagmentation fragments. The 2nd sequencing read primer region is attached to the 3´ ends later in the protocol. Thus, all tagmentation fragments are treated in the same way at all steps. This approach is different to the Illumina Nextera protocol where transposomes bear randomly paired transposon end adapters of two types – type A with the 1st and type B with the 2nd sequencing read primers regions. As a consequence, half of the library molecules have the same flanking sequences and do not participate in amplification [Adey et al., 2010; Caruccio et al., 2011]. Sequential introduction of two types of Illumina technical sequences is one possible way to exclude such a loss of tagmentation fragments. In the Discussion we describe alternative transposon structures, providing symmetrical ligation of technical sequences to the fragment ends.

Another difference to the Illumina transposon structure is that the mosaic end (ME) part of the transposon end adapter has a shorter double-stranded part – 15bp instead of 19bp. Decrease in length facilitates the oligonucleotide replacement in the subsequent gap repair step.

Being dependent on the presence of Mg^{2+}, tagmentation reaction is stopped by adding EDTA to exclude any residual tagmention during sampling of an aliquot for further processing.

Dilution of the tagmentation reaction

During this step the actual PTULI library preparation starts. Based on calculations given in the example above, it is determined which amount of DNA should be taken for the library preparation. For sequencing on an Illumina sequencing lane - about 1/1000 of the tagmentation reaction needs to be processed further. Manipulating DNA at such low concentrations has its risks, for example, DNA can stick to the tube walls or get significantly lost under standard pipetting. To avoid this problem we use 2ng/µl tRNA as a dilution carrier for the DNA sample. tRNA yeast was demonstrated to be inert for DNA applications and not to inhibit amplification [Ruijter et al., 2013]. Our internal test confirmed that tRNA does not give rise to any library preparation artifact products (Figure 12).

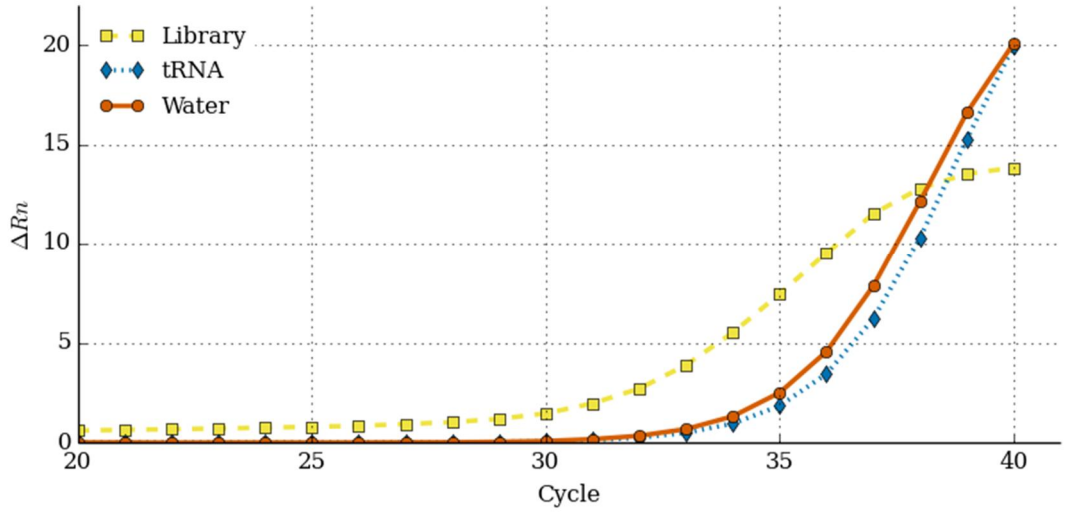


Figure 12 Evaluation of tRNA effect on PTULI library preparation. Tagmented DNA diluted in 2ng/µl tRNA underwent Tn5 removal and gap-repair steps of the PITS library preparation protocol along with 2ng/µl tRNA and water as negative controls. Subsequent qPCR shows similar later rise of the water and tRNA amplification plots relative to the actual library plot.

Gap repair reaction

In an aliquot of tagmentation reaction with preserved contiguity and calculated amount of molecules, Tn5 transposase can be removed, because all arising tagmentation fragments are intended to be processed and sequenced.

The transposase inserts breaks into the strands of dsDNA at 9nt apart from each other. Each strand of the tagmentation fragment gets ligated to the transposon's Tn5ME-A oligonucleotide on the 5´ end. Both 3´ ends remain recessive, with a 9nt distance to the 5´ end of the transposon's p_026 oligonucleotide, which is not covalently linked to the tagmented DNA (Figure 11). In the PTULI protocol p_026 is replaced with the p_028 oligonucleotide, containing the 2nd sequencing read primer region in the 3' part (marked blue in Figure 11).

The 9nt gap is closed in a gap repair reaction. A gap repair reaction (*in vitro*) is known for its specificity and has been successfully used in SNP detection over a decade [Hardenbol et al., 2003]. Normally, a combined action of a polymerase extending the 3´ end and a ligase sealing the nick is exploited in this technique. If a 5´ end of the template is not phosphorylated, a T4 PNK can be added to the mixture. A critical requirement to polymerase is that it should lack a strand displacement activity [Sambrook and Russell, 2001].

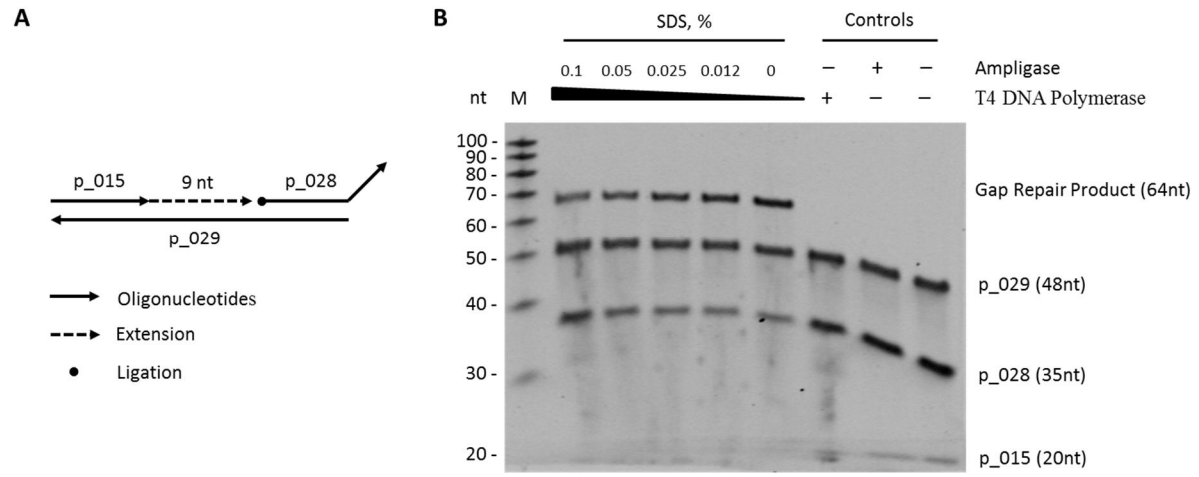


Figure 13 (A) Oligonucleotide model used for gap repair reaction setup. Oligonucleotides p_015 and p_028 are annealed to p_029 forming a 9nt gap. (B) Influence of SDS concentration on the gap repair reaction. Yield of the gap repair product in the presence of 1:2 serial dilutions of SDS was monitored on 10% PAGE. Marker: Ladder 20/100 (M).

The gap repair reaction for the PTULI procedure was first tested on a model oligonucleotide system (Figure 13A). T4 DNA polymerase was tested in tandems with T4 DNA ligase and thermostable Ampligase (Supplementary Figure 1). Though in both cases the reaction yield was quantitative, we chose Ampligase. This is because a non-specific ligation product was observed in the absence of polymerase, when using T4 DNA ligase (marked with an asterisk in Supplementary Figure 1B). A non-template ligation has also been previously reported for T4 DNA ligase (Kuhn and Frank-Kamenetskii, 2005). Using test oligonucleotides and monitoring the gap repair product outcome, the SDS concentration for Tn5 removal was determined (Figure 13B). To avoid purification step before the gap repair, it was essential to keep the SDS concentration as low as possible not to compromise the efficiency of the gap repair reaction, but still within the working SDS concentration for Tn5-based protocols - 0.01-0.2% [Picelli et al., 2014; Amini et al., 2014].

Oligonucleotides modifications were used as in [Wang et al., 2013]. The transposon's bottom oligonucleotide p_026 bears a dideoxycytidylate at 3´ end to prevent unwanted extension during the gap-filling reaction and, later, during PCR [Sram et al., 2008]. It is also shorter than used in the standard tagmentation NGS protocols, to facilitate dissociation at lower temperatures and avoid denaturation of the tagmentation fragments. We tested another modification of the bottom primer containing internal uridines (p_027) to use UDGase for its removal. However, this modification interfered with the tagmentation performance (Figure 22, explained in the 'FEA' part of the Results section). The replacement p_028 oligonucleotide has an inverted deoxythymidilate at the 3´ end, leading to 3´-3´ linkage,

which inhibits both degradation by 3´ exonuclease activity of T4 DNA polymerase and unwanted extension by DNA polymerases [Shaw et al., 1991; Seliger et al., 1991; Ortigao et al., 1992].

As a result of the gap repair step, all 3´ ends of the tagmentation fragments acquire sequence corresponding to the 2nd sequencing read primer region. As is seen on Figure 11, tagmentation fragments are now flanked with symmetrical Y-shaped adapter sequences. This ensures that all tagmentation fragments have the required flanking regions to be amplified. Moreover, each fragment is represented in the amplification as two templates, since each strand can be amplified. This feature secures the protocol and increases the probability of each fragment to be represented in the final library.

Amplification and loading on a flowcell

For standard sequencing applications like RNA-Seq or exome sequencing only a small part of the NGS library is loaded on a flowcell. For example 50nM library in 30µl is a typical result for mRNA library preparation starting from 1µg of total RNA. From this amount (1500fmol) only 20fmol are taken for the denaturing of the library, and roughly 120µl x 10pM = 1.2fmol are loaded on the HiSeq2500old flowcell, where 120µl is the loading volume and 10pM is an approximate recommended loading concentration for Illumina platforms with non-patterned flowcells. The difference between total and sequenced amount of the library is 3 orders of magnitude, which correspond to 10 amplification cycles. In the case of the PTULI library, such waste of material is unacceptable. For the low input libraries the more amplification cycles are performed the larger is the distortion of the proportion of the amplicons and the probability of artifact products is higher. Besides, in a small aliquot taken for sequencing, purely statistically, not all types of molecules might be present.

We worked out a procedure for non-residual loading of PTULI libraries on a sequencer. The total amount of required library molecules is determined by the sequencing platform specifications, - for example, 6 fmol for Illumina MiSeq, counted for 10pM of the recommended loading concentration and 600µl of the final loading volume. The volume of the library is dictated by Illumina flowcell loading procedure, where the library is first denatured in 0.1N NaOH and then neutralized and at the same time diluted to the loading concentration in the Illumina HT1 buffer. It is important to keep the final NaOH concentration as in the Illumina loading protocol ($\leq$ 0.01N), because the change of this parameter might be unpredictable for surface amplification [Quail et al., 2008]. Within the 600µl of the loading volume, the maximum possible volume of denatured library in 0.1N NaOH is only 6µl. For

PTULI libraries we aim at 5µl of library volume leaving 1µl for 0.6N NaOH (see Materials and Methods section). This means that ~ 15x 10^6 tagmentation products in 10µl of the gap repair reaction have to be amplified to reach 6 fmol and concentrated in 5µl. Since amplification can be performed only once, several **test amplification (TA) libraries** are prepared from equal aliquots of tagmented DNA in parallel to the main library. For concentration measurement a **reference library** is used - a PTULI library, which was amplified to achieve a concentration measurable by Qubit and sequenced to determine the optimal loading concentration. After the gap repair, one of the TA libraries is amplified, purified, diluted in 5µl and its concentration is measured in qPCR relative to the reference library. Based on the concentration measured, the number of cycles is adjusted (as in the example in the Supplementary Figure 2) and tried on the second and, if necessary, on the third aliquot. The selected number of cycles is then applied to the main sample. The results obtained on TA libraries proved to be reproducible, and though processing of the main sample is blind, it works reliably. The required amount of library for loading may be optimized basing on the known loading concentration of the sequenced reference library. It is important to note that while reference library can originate from other tagmentation reaction than that of the main library, TA libraries must be prepared from the same tagmentation reaction as the main library.

The whole gap repair reaction goes into amplification. To minimize inhibition effect of the reaction components on amplification, PCR is performed in a 20 times larger volume than the gap repair. Qiagen PCR purification columns are used to clean up and concentrate the amplification reaction to most fully preserve the existing fragments size range. Elution volume is minimized by using MinElute columns and brought to the desired 5µl volume through evaporation.

Recommended experimental setup

In the above description and discussion of the individual steps of the PTULI protocol several controls, necessary to monitor the efficiency and specificity of enzymatic reactions and to select the required amount of the amplification cycles, are mentioned. Figure 14 gives an overview of the optimal, in our view, experimental setup for the PTULI library preparation. When working with minute amounts of input DNA, extreme care should be taken to minimize the losses and to avoid contamination with other samples and amplified material. Some steps of the PTULI protocol involve nucleic acid molecules being present at higher concentrations than the input genomic DNA, for example carrier tRNA, adapter oligonucleotides, PCR

primers, which can cause the formation of artifact products. Therefore, along with the main sample, we recommend to run several control samples and reactions in parallel, which would either reveal the problem(s) or serve as a proof of the proper performance of the protocol.

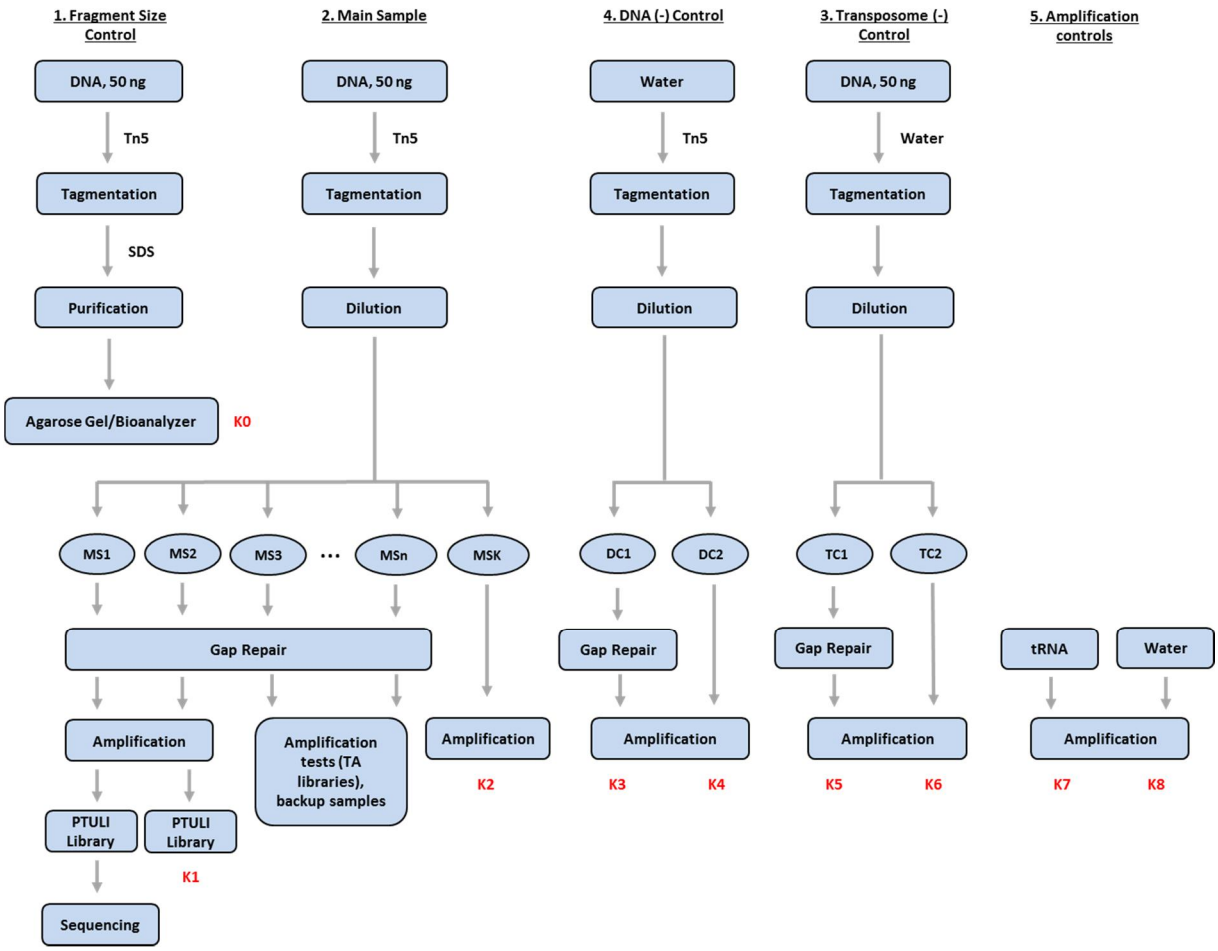


Figure 14 Overview of the recommended experimental setup of the PTULI library preparation procedure.

A size control sample (1) is added to the protocol to monitor fragmentation efficiency (K0). For the Main Sample (2) several identical aliquots of diluted tagmentation reaction are prepared (MS1-n, MSK). MS1 is the aliquot for preparation of the PTULI library which is going to be sequenced. One of the aliquots (MS2) is processed exactly as the MS1 and gives the possibility to check library molecule size distribution and eventual sequencing adapters contamination before loading the MS1 library on the sequencer (K1). Some aliquots are used as test amplification (TA) libraries to determine the number of amplification cycles. MSK is an amplification specificity control – it ensures that no unspecific or contamination products are amplified in the sample if the second PCR primer sequence is not introduced during the gap-repair reaction (K2). MS aliquots require no additional input material, as the initial

tagmentation step is performed on a huge excess of material relative to the amounts used for the library preparation.

DNA template negative (3) and Transposome negative controls (4) are necessary to exclude or to reveal contamination of the reagents with ready library molecules and to confirm that no by-products are formed by participating oligonucleotides (K3-K6).

Amplification controls are standard for any amplification experiment and monitor the purity of the amplification reagents (K7 and K8).

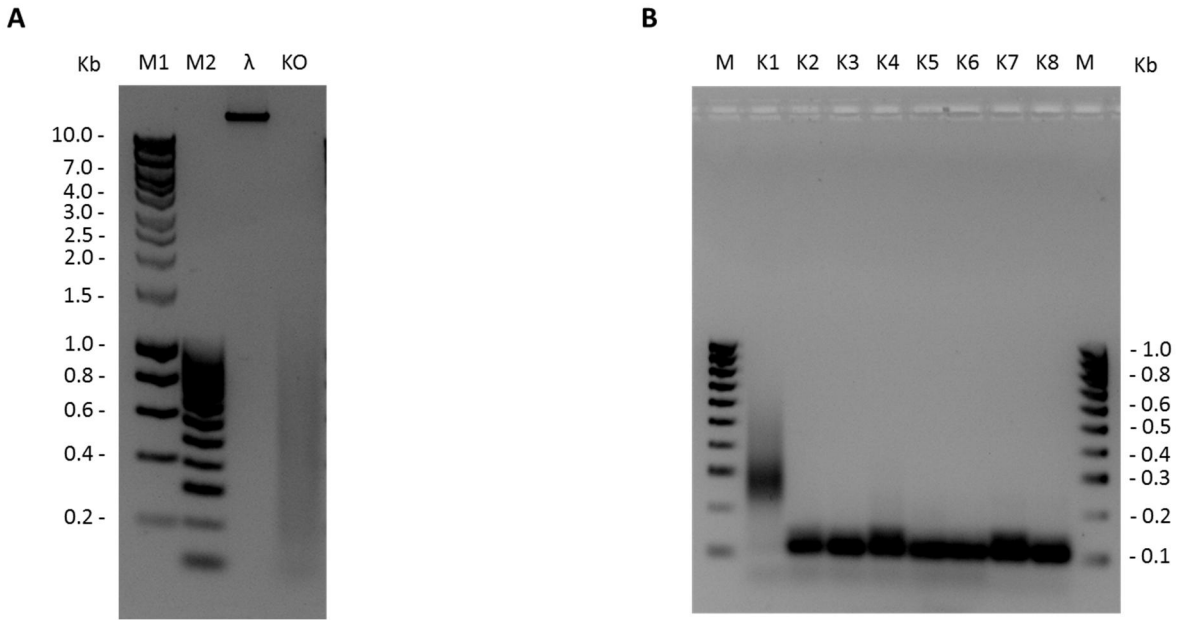


Figure 15 Analysis of PTULI library preparation performance in accordance with the experimental setup depicted in Figure 12. (A) Size control of the tagmentation reaction. 50ng of phage λ DNA was tagmented with 16µl of Tn5ME-A transposome assembly in total reaction volume of 5µl. (B) PTULI library size distribution and library preparation controls. K0-8 correspond to the controls explained in Figure 12**.** K0 – tagmented phague λ DNA: PTULI ready library (K1); λ DNA, Tsomes, gap-filling(-) (K2); λ DNA(-), Tsomes , gap-filling (K3); λ DNA(-), Tsomes, gap-filling(-) (K4); λ DNA, Tsomes (-), gap-filling (K5); λ DNA, Tsomes (-), gap-filling (-) (K6); tRNA, 2ng/µL (K7); water (K8). HyperLadder 1kb (M1), HyperLadder 100bp (M2) Amplification was performed with primers p_032 and p_033, 40 cycles. 1.5% agarose gel was run for 1h 30min, 150V.

Control reactions can be analyzed on an agarose gel or Bioanalyzer. Figure 15 shows an example of such analysis for a PTULI library prepared from phage λ DNA. Lanes of the agarose gel in Figure 15 are signed with K0-8, the numbers corresponding to the control numbers in Figure 14. Except for distribution of the tagmented DNA sizes, other controls can be also monitored during amplification. Real-time PCR plots for K1-8 can be seen in Figure 16.

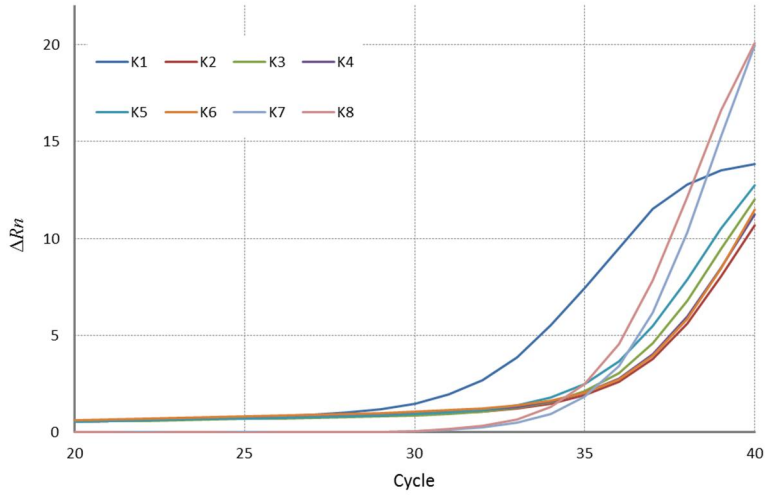


Figure 16 qPCR analysis of the PTULI library preparation performance in accordance with the experimental setup depicted in Figure 14. K0-8 correspond to the numbering used in Figures 14 and 15.

The PTULI library illustrating experimental design of the protocol is prepared from only ~90000 molecules of phage λ (= 0.16amol, or 5pg of DNA). Gel image confirms the library has no visible contamination and all controls look as they are supposed to – empty. In the next section sequencing data obtained from this library will be discussed.

Proof-of principle paired indexing experiment

Indexing scheme

The PTULI protocol was developed for implementation of the PITS approach. Figure 11 does not show the location of tagmentation sites indices to avoid the distraction from the library preparation procedure. Paired indices may be integrated in different transposons structures (Figure 10). It is particularly important that the tranposon strands containing the index sequence get covalently linked to the 5´ ends of the tagmentation fragments during the tagmentation step, where the contiguity of sequence and the pairing of indices are unquestionable.

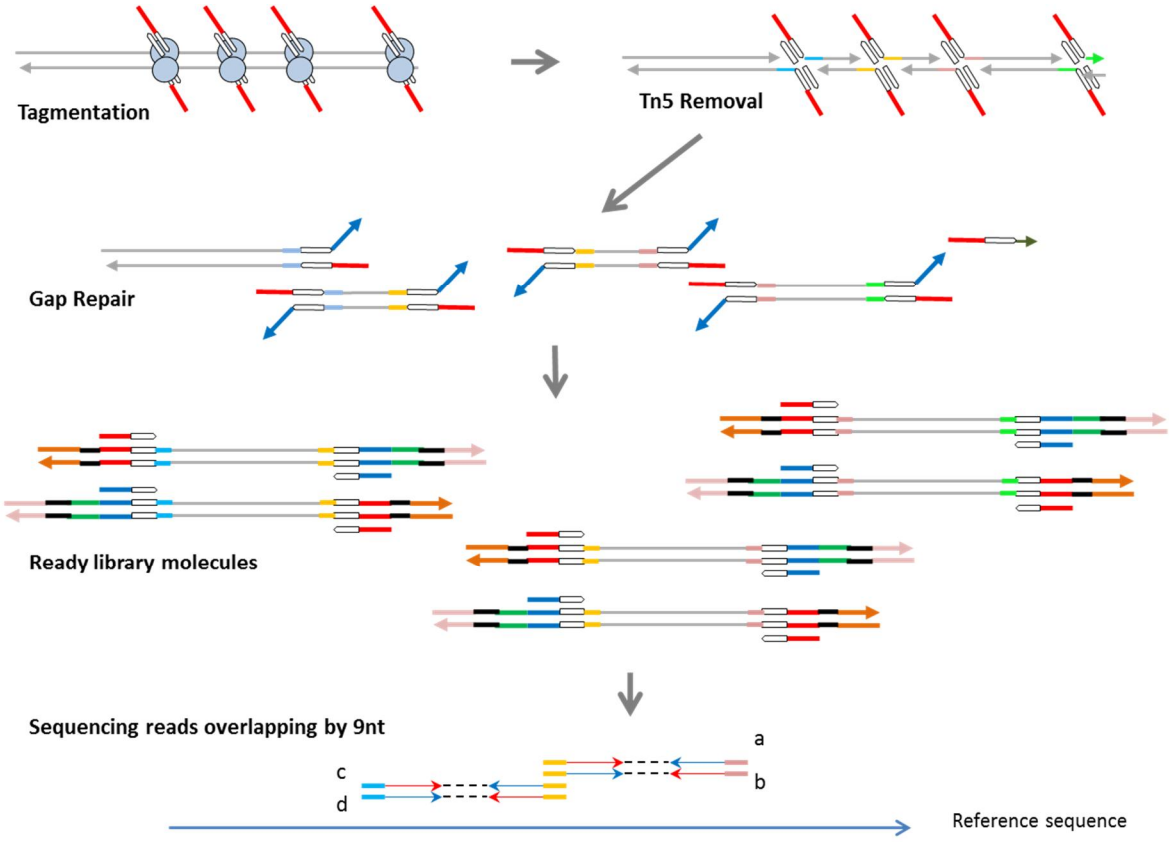


Figure 17 Natural paired indexes at tagmentation sites. Tn5 transposase cuts the strands of the double-stranded substrate with a 9nt asymmetry. Complementary 9nt stretches get split between two neighbor fragments and preserve in the library molecules. Sequencing reads overlapping by 9nt point to the same tagmentation site. Sequenced fragments a, b, c and d are shown as read pairs (1st read - red arrow, 2nd read - blue arrow, dashed line – eventually missing sequence).

For the initial experiments, it was desirable to avoid ordering multiple oligonucleotides with index sequences and use as simple a scheme as possible. And here again the very nature of the transposase action prompted a reasonable solution. 9nt long recessive 3´ ends of tagmentation fragments may serve as codes: they randomly emerge in the transposition process and are naturally paired as complementary sequences by origin. Since these sequences are not lost further in the PTULI protocol but the second strand is restored for each of them, they remain part of sequencing libraries inserts. Figure 17 explains this particular case of paired indices. 9nt stretches next to ME regions in the tagmentation fragments marked with the same colour belong to the neighbouring fragments. The reads from these fragments would overlap by 9nt.

These ‘naturally paired indices’ would be most advantageous in combination with the paired ‘artificial’ indices in transposons adding to the diversity of indices the diversity of indices combinations. They could be also used to determine or verify pairs of `artificial´ indices in case the later are not identical or complementary sequences. Used alone `naturally paired indices´ are inappropriate for reconstruction of large genomes and repetitive sequences,

because the 9nt overlaps would not be unique. However, they are sufficient for resequencing of small non-repetitive templates.

Template selection

Phage λ genomic DNA was selected as a test tagmentation template for the PITS and PTULI protocols, mainly because it is convenient to deal with a countable number of genomes of fixed length in test experiments. Phage λ genomic DNA is available commercially, is of standardized quality and full length.

Use of the phage λ DNA in the PTULI library preparation protocol is advantageous because it represents a case if an imperfect input material, where the resulting λ PTULI library is of a low-complexity. If a protocol, aimed at low input material, works with a low complexity DNA, where there is more probability that some region would get overamplified, it would also work on a more complex material.

Finally, for the paired indexing scheme we needed a well-known genome for univocal mapping of sequencing reads. Short size of the genome was also thought to compensate for the loss of the fragments during tagmentation through using more copies. This would allow to tune paired indexing independently of single-molecule assembly.

In all our preparation tests we started with 50ng of DNA Lambda phage. This amount is easy to handle and is sufficient for a smear visualization.

PITS libraries sequencing

Two PITS libraries, prepared from different amounts of the starting material, were sequenced on Illumina MiSeq platform. Pre-sequencing information about these two libraries is given in Table 6.

Table 6 PITS libraries taken for sequencing

Library	Number of cycles	Volume, µl	ng/µl	Total fmol*	Loaded on MiSeq, fmol*
Library90000	27	20	8.5 (Qubit)	1005	6
Library4500	24	5	0.97(qPCR)	6	6

*Counted for 260bp average library molecule size

Library90000 was planned to be loaded on a HiSeq2500old flowcell lane, so according to the theoretical estimate of 300bp (average expected size of the fragments) x 15 x 10^6 / 48502bp (size of the phage λ genome) = ~90000 molecules of phage λ were taken from the tagmentation reaction for PTULI library prep reaction. Library90000 served as a Reference library in PTULI experiments, so it was amplified to the amounts measurable with Qubit.

Library4500 was planned to be loaded on a MiSeq flowcell. According to the reference library, base estimate of 130bp (average size of the fragments in the reference library) x 1.5 x 10^6 / 48502 bp (size of the phage λ genome) = ~4000 molecules of phage λ should have been taken from the tagmentation reaction. We decided to start with 4500 phage genomes, because we aimed at larger size of library molecules.

This library was prepared to be loaded completely on a sequencing flowcell, so the number of amplification cycles and concentration was determined according to the reference Library90000. Amplification was quite predictable: the difference in the total amount of ready library molecules (167 times) was explained by the difference in the starting amount (20 times) and larger amplification range (3 cycles – 8 times difference).

Sequencing statistics for both libraries are presented in Table 7.

Table 7 Sequencing statistics for the two sequenced libraries

Characteristics	**Library90000**	**Library4500**
Total reads	15581524	11620488
Mapped (% of total)	15477770 (99.3%)	11522343 (99.2%)
Non-duplicated, properly paired reads (% of total)	176838 (1.13%)	16878 (0.14%)
Insert size (mean/median)*	130/157.6	190/205.5
Average coverage (length of phage λ genome = 48502bp)*	120x	11x
Unique start sites*	28668	8443

* counted for the non-duplicated, properly paired reads

The two libraries were prepared from independent tagmentation reactions. A smaller number of reads for the Library4500 is explained by the larger insert sizes (Figure 18A): 190bp insert size means that library molecules were around 320bp (sequencing adapters flanking the library molecules comprise 129bp), so the actual loaded amount was not 6fmol but 4,86fmol, and there should be ~1.24 times less clusters. We obtained 1.34 times less clusters, which is very close.

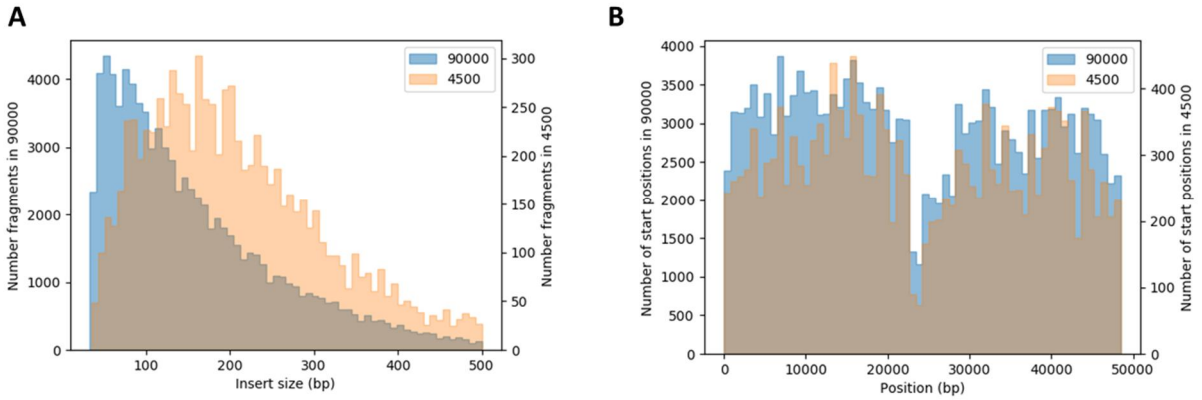


Figure 18 Sequencing results for Library90000 and Library4500. (A) Insert size distribution in the libraries. (B) Distribution of reads start positions along the genome of phage λ.

Both libraries showed high percent of mapped reads and no contaminating adapter sequences artifacts, which often compromise the results of sequencing of low complexity libraries. Figure 18B shows the mapped reads distribution along the phage genome. High average coverage also confirms the high quality of the libraries. Distribution of the start sites along the lambda genome was similar for both libraries. It was not quite even due to the transposase preferences but reproducible. About one third of all positions in the lambda genome were not tagmented in our experiments. Supplementary Table 1 shows 25 most and 25 less represented tagmentation sites in the phage lambda genome selected from the sequencing data for both libraries.

The difference in the number of unique reads and coverage between the two libraries was ten times and not 20, as the difference in starting material would suggest. This can be explained by the higher probability of getting identical tagmentation fragments from a larger number of genome copies in Library90000. Without external indices we cannot discriminate between identical tagmentation products, originating from different genome copies, and PCR replicates of the same tagmentation product. When increasing the number of tagmentation template molecules, there is a point when the number of non-duplicated reads would stop growing. This stagnation is influenced by transposase preferences.

For the number of genome copies used for test libraries preparation (90000 and 4500) the probability of occurrence of tagmentation sites at the same position within the different genome copies is quite high. In an ideal case, if tagmentation occurs truly randomly, the probability of tagmentation at each genomic position is roughly 1/50000 = 2 x 10^{-5}, counting phage λ genome length is ~50000bp. For 250bp average tagmentation fragment size or ~200 tagmentation events per phage λ genome copy the probability to cut the same position in two different genome copies is p = 1 – $(1 – 2 \times 10^{-5})^{200}$ = 0.004. If we start with *n* genome copies, the probability to get *k* tagmentation events at the same position corresponds to the binomial distribution ~ B (n, p). For n = 4500, p = 0.004 the probability that k>1 is 0.999998, and practically none of the tagmentation sites in the test libraries should be unique. Around 200-300 phage λ genomic copies at the start of PITS library preparation are required for univocal reconstruction of scaffolds consisting of ≥ 5 fragments. In the reality situation might be different. Tagmentation sites cannot be too close to each other, Tn5 has a sequence preference and also some fragments will be lost in the course of library preparation. With this background it was interesting to analyze the obtained sequencing data.

We tried to identify the neighbouring fragments, sharing the same 9nt of the paired index, within the non-duplicated, properly paired reads. The numbers obtained so far are summarized in Table 8.

Table 8 Paired indexing statistics

Characteristics	**Library90000**	**Library4500**
Library inserts with unique start (first 9nt of the 1st read) and end (first 9nt of the second read)	88251	8429
Library inserts harbouring paired indices which could be joined in a scaffold	76310	4687
Unique scaffolds: total	56	104
Unique scaffolds: with 2 library inserts	54	99
Unique scaffolds: with 3 library inserts	2	4
Unique scaffolds: with 4 library inserts	0	1

In a unique way, it was possible to restore quite a few scaffolds – 56 for Library90000 and 104 for Library4500. The tendency – the increasing number and length of scaffolds with the decreasing number of starting material - corresponds with the expectation. Now that it is known that the PTULI protocols works, it is possible to lower the amount of initial DNA material.

Currently, the same 9nt indices are shared by more than two library inserts. For example, from the four fragments (1) A- B, (2) B- C, (3) C-D and (4) C–E, where letters correspond to 9nt indices, it is possible to reconstruct a chain of 2 fragments: (1) A-B + (2) B-C. The further choice is ambiguous. For most of the fragments sharing 9nt index sequence (Table 8), this occurs already at the first step. If the scaffolds are not restored in a unique way, they will be much longer. Currently, we are developing an assembly algorithm to gather such scaffolds and simulate *de novo* assembly process for phage λ DNA.

Preparation of in-house Tn5 Transposase

Tn5 Transposase expression and purification

The pTXB1-Tn5 vector was designed using the IMPACT system (NEB) which leverages the inducible self-cleavage activity of protein splicing elements to separate the target protein from the affinity tag. The Tn5 protein was fused at its C-terminus to an intein tag (~28kDa), containing the chitin binding domain (6kDa) (Figure 19A).

Tn5 Transposase expression was performed as described [Picelli et al., 2014] with insignificant differences in technical performances at several stages. The protocol has proved that cloning in the pTBX1 vector is more advantageous in comparison with the pTYB4 vector [Bhasin et al. 1999] and two mutations only (E54K and L372P) are responsible for Tn5 hyperactivity [Picelli et al., 2014]. The efficiency of each stage of Tn5 production protocol was controlled by loading control aliquots on SDS-PAGE gel (Figure 19B).

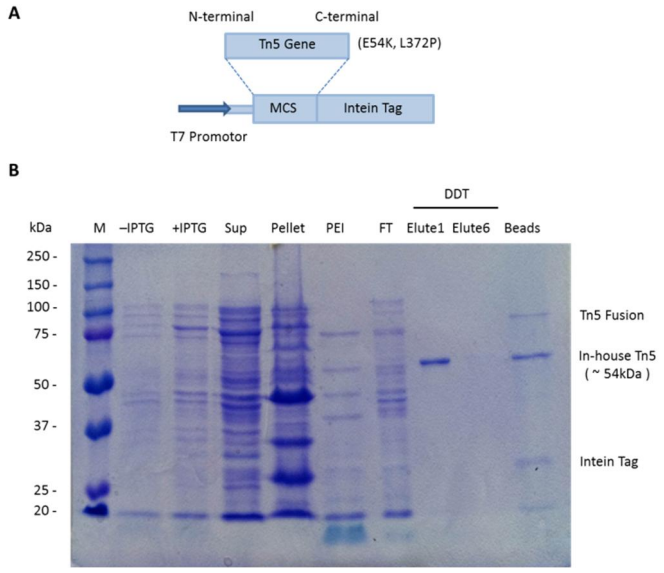


Figure 19 Tn5 cloning scheme (A). Production of In-house Tn5 Transposase (B). 10% SDS-PAGE run with crude extract from uninduced cells (–IPTG), crude extract from cells induced at 25°C for 4h (+IPTG), clarified supernatant (Sup), pellet (Pellet), PEI clarified supernatant (PEI), flowthrough from chitin beads (FT), elution (1^{th} fraction) of cleaved protein (Elute1), elution (6^{th} fraction) of cleaved protein (Elute6), chitin beads aliquot after elution (Beads). Marker: Precision Plus Protein Dual Color Standard (M).

Approximately 1ml of stock Tn5 solution at a concentration of 41µg/µl (about 750pmol/µl) was obtained from 1l of induced bacterial culture. We obtained an adapter-free Tn5 In-house Transposase and used different transposome assemblies depending on the experiment. To note, additional washing fractions of chitin-beads could be included in the protocol to increase the Tn5 yield which is confirmed by a control aliquot: there still enough protein linked to beads (Figure 19B, line Beads).

Transposase activity assay

We assessed the activity of the assembled Tn5 through its ability to tagment linearized pKSII plasmid DNA (Supplementary Figure 4A). Four titration points were used to illustrate the tagmentation efficiency by volume titration of Tn5 Transposome mixtures. Tagmentation efficiency is characterized by the average size of DNA fragments which decreases with higher volumes of Transposomes. As a tagmentation control, we performed one reaction with Tn5 protein from our colleagues. The concentrations of the two Tn5 batches were adjusted to work within a similar range. Normally, all of the pKSII plasmid DNA was converted to fragments. An image of a typical activity assay is shown in Figure 20.

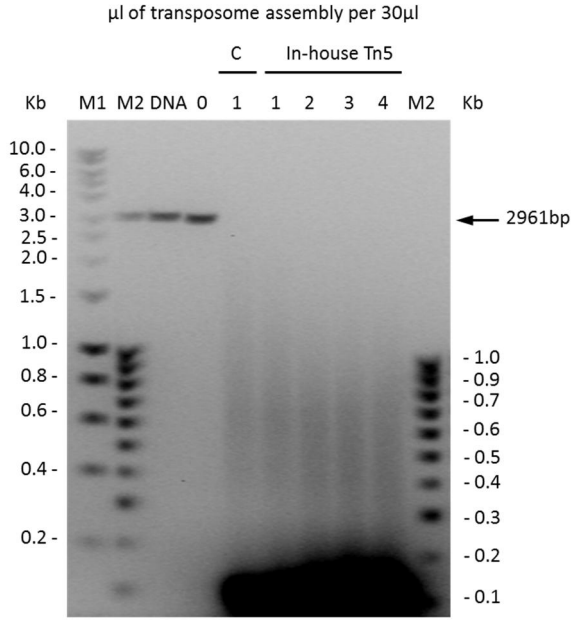


Figure 20 In-house Tn5 Transposase activity assay. Transposomes were assebled with Tn5ME-A adapter. Tagmentation reactions were performed for 30min at 25°C with 50ng of linearized plasmid pKSII/EcoRI. The reactions were stopped by adding 2% SDS to final concentration of 0.08% and heating for 7mins at 55°C. The resulting products were visualized on 1.2% agarose gel (100V for 4h) without purification. Markers: HyperLadder 1kb (M1), HyperLadder 100bp (M2). Samples: linearized plasmid pKSII/EcoRI (DNA), control transposomes prepared with characterized Tn5 protein (C).

Fragmentation efficiency assay (FEA)

Tagmentation is a key step of the PITS approach. It is crucial to have a convenient instrument to estimate tagmentation efficiency. Batches of Tn5 tranposase and transposomes need to be compared to provide the same experimental conditions and, thus, consistency of tagmentation. Optimal transposome assembly and tagmentation reaction settings and conditions have to be determined, for example, reaction duration, reaction buffers, transposon structures etc.

At the outset of PITS protocol development, for instance, we aimed at obtaining maximally active Tn5 transposomes. More active transposomes would provide shorter NA fragments and facilitate an optimal library size which is highly important to meet sequencing requirements. The size of an insertion flanked by adapters is determined by the limitations of the NGS instrumentation. When using Illumina technology (MiSeq), the optimal library length is

restricted by the cluster generation process which is relatively inefficient. Although there are examples of successful sequenced libraries with insertions up to 1500bp, current Illumina protocols imply to have a fragment size varied in the range of 200-500bp [Head et al., 2014]. In the case of larger fragments (up to 1000bp), clusters will still be generated but with increasingly lower efficiency and yield, which should be taken into consideration [Bronner et al., 2009].

For the further PITS related technology development, design of the transposon might need to be optimized to exclude the gap repair step. It is also essential to monitor home-made Tn5 transposomes stability to work under the same experimental conditions.

At the time of conducting our research the only available option to evaluate the efficiency of transposome assemblies was a visualization of DNA smear on a gel. This approach is tedious and, in many cases, inappropriate for technology development applications. In the beginning of our work with Tn5 enzyme we found that evaluation of the smear on a gel was not sensitive enough for a number of experiments. This problem has been already raised by several authors. *Bogdanoff et al.* attempted to characterize the efficiency of Tn5 transposomes using comparative qPCR but their approach failed methodologically emphasizing a demand for a working technique.

It was critical for us to have a convenient non-electrophoresis method for estimation of tagmentation efficiency. We developed a qPCR based DNA fragmentation efficiency assay (FEA) to characterize the performance of Tn5 transposomes. The assay was published and all the further details relating to it are in the accompanying paper [Rykalina et al., 2017]. Here we describe the general idea of our the method and provide some examples of using it for the work related to the PITS protocol.

The general principle of the assay is shown in Figure 21. Tagmentation reaction in which DNA molecules are fragmented and tagged with the adapters, is characterized by performance of Tn5 transposomes. In our assay, we use a plasmid DNA as a reference substrate for tagmentation reaction (Supplementary Figure 4B). Efficiency of Tn5 transposome is analyzed by comparative qPCR by detecting the difference (DCt) in amplification of the certain plasmid regions before and after tagmentation. In the case of a more efficient fragmentation, we observe a larger number of cleavage events within an amplified region, which gives a delayed raise of the amplification curve and, respectively, a larger DCt. In addition to the main idea of the assay, we also introduce a reference tagmentation template as a spike-in which can be easily added to the target DNA to monitor its fragmentation efficiency along a library preparation procedure.

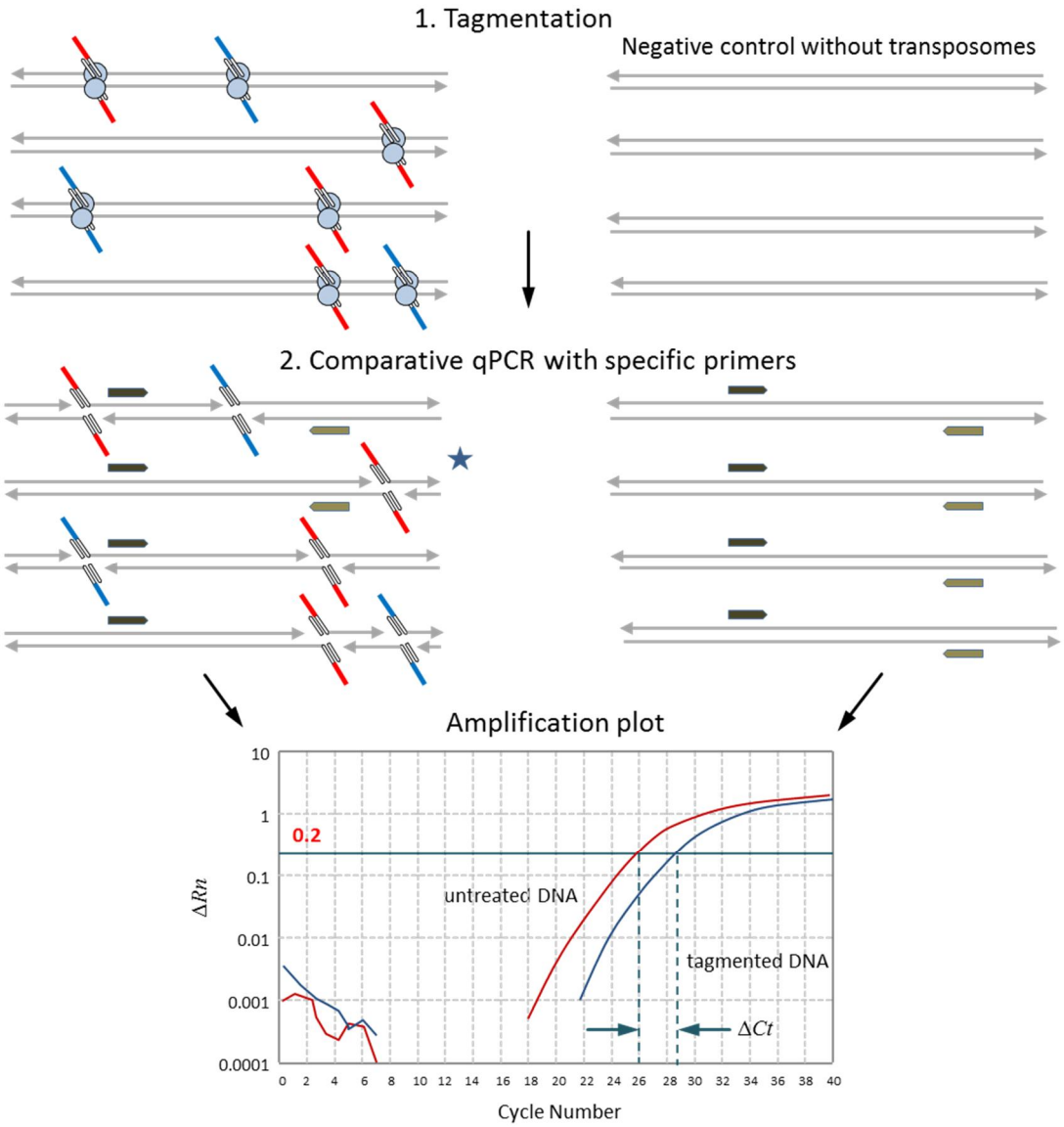


Figure 21 Schematic view of the principle of the Tn5 transposomes fragmentation efficiency test. During the tagmentation (step 1) equal amounts of plasmid molecules are processed in the same conditions in parallel with (left column) and without (right column) transposomes. Transposomes are depicted as double circles, each circle with partly double arrows, corresponding to transposase dimers bound to Illumina oligonucleotide adapters. Transposase recognition sites are shown as empty double arrows and two types of single stranded tails are colored with yellow and blue. After removal of transposomes, samples are analyzed with real-time PCR (step 2). PCR primers are plasmid-specific and shown as green and orange arrows. All molecules in the untreated sample can be amplified. In the transposome-treated sample only those molecules may be amplified which have no transposase-inserted breaks in the region between the PCR primers: from the four drawn DNA molecules only one (marked with a star) gives rise to a PCR product. The amplification curve demonstrates the difference in Ct (here two cycles) corresponding to the difference in the amount of amplifiable templates in tagmented and untreated DNA samples (here four times) [Rykalina et al., 2017].

Described below are several PITS related parameters which were evaluated by FEA and in parallel with agarose gel analysis.

In the current PITS library preparation procedure, we use the Illumina Tn5ME-A transposon end adapter only. The sequence of this adapter differs slightly from a standard Illumina

adapter by the size (4nt shorter for a bottom part) and modifications. Tn5-based library construction with only one adapter was first introduced for a bisulfite sequencing protocol [Wang et al., 2013]. The authors showed that such an adapter performs well in a tagmentation reaction followed by an oligonucleotide replacement and gap repair.

By FEA we compared the extent of NA fragmentation with Tn5 transposomes assembled with a modified Tn5ME-A adapter (p_026 bottom oligonucleotide) along with an adapter of the same structure, but with 3 uridines replacing thymidine in its sequence (p_027 bottom oligonucleotide). Our results proved tagmentation efficiency of the A type adapter suggested by *Wang et al.* and complete inapplicability of the one with uridines (Figure 22). From the one hand, this particular example illustrates yes/no interpretation which could be theoretically analyzed by an agarose gel. From the other hand, the influence of uridines on fragmentation efficiency was not that obvious. In the case of a less dramatic picture, the sensitivity of the gel visualization might not be enough. This finding is very important in the context of other possible adapter schemes, including lampion-like and double-stranded structures.

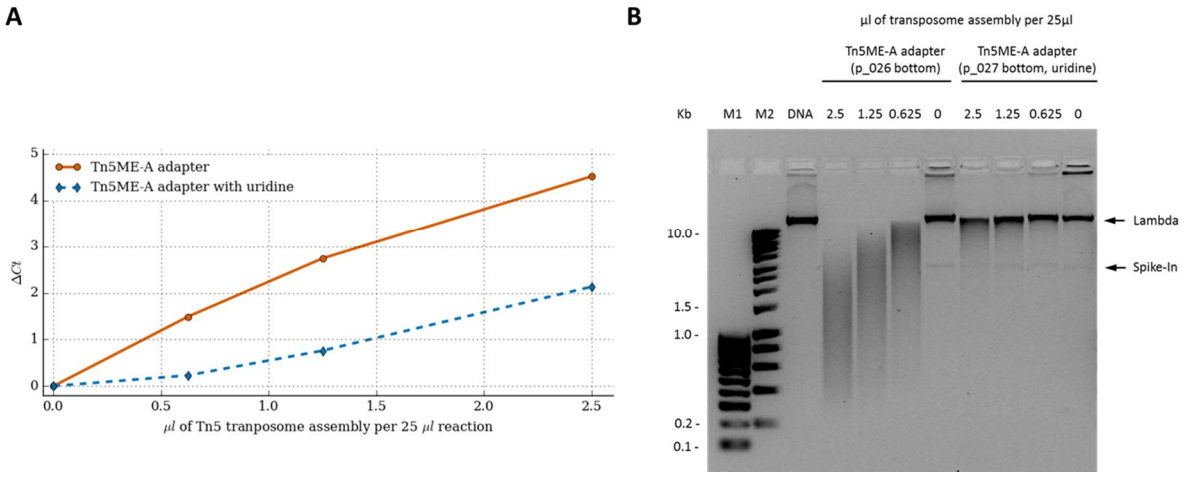


Figure 22 Comparison of adapter structures. 1.25μM Tn5 transposome assemblies were prepared by incubating Tn5 enzyme with two types of Tn5ME adapter for 30 min at 25°C. 50ng of Lambda DNA was tagmented with a reference tagmentation template (pUC19/EcoRI) for 8min at 55°C. The reactions were stopped by adding 2% SDS to final concentration of 0.16% and heating for 7mins at 55°C. The resulting mixtures were purified with AMPure DNA beads and analyzed by (A) FEA or (B) loading on 1.1% agarose gel. Markers Markers: HyperLadder 1kb (M1), HyperLadder 100bp (M2).

Our Tn5 transposomes are not preassembled with adapters of a certain type and it takes advantage of varying conditions of the protocol. The disadvantage of small scale purpose-specific preparation of transposome assemblies is an excess of unbound adapters. For most standard applications, this is not an issue because free adapters can be removed by simple purification on beads. However, there still might be an application where unbound adapters inhibit a reaction. In the FEA experiments we used beads to get rid of free adapters as their

excess affects qPCR. Analogously, overall PITS library preparation protocol can be depressed by a high concentration of adapters in the tagmentation reaction owing to high working volume of transposomes. We tried to purify resulting transposome assemblies from an excess of transposons using filter units with 100K pore size aiming at capturing Tn5 homodimer complex. In principle, the filtration on Amicon units is a simple procedure but it still needs some improvements: transposome solution with a high concentration of glycerol is harder to centrifuge. As one can see in Figure 23, Tn5 transposome assembly after filtration manifests a similar fragmentation efficiency in comparison with the intact assembly. Although *Picelli et al.* demonstrated that Tn5 adapters can be annealed to Tn5 already on the column, during protein purification they did not show accurate tests comparing activity of the transposomes under the same working volumes [Picelli et al., 2014]. To our knowledge longer incubation periods of transposome assemblies under increased temperature (36-48h at 4°C) adversely affect their activity.

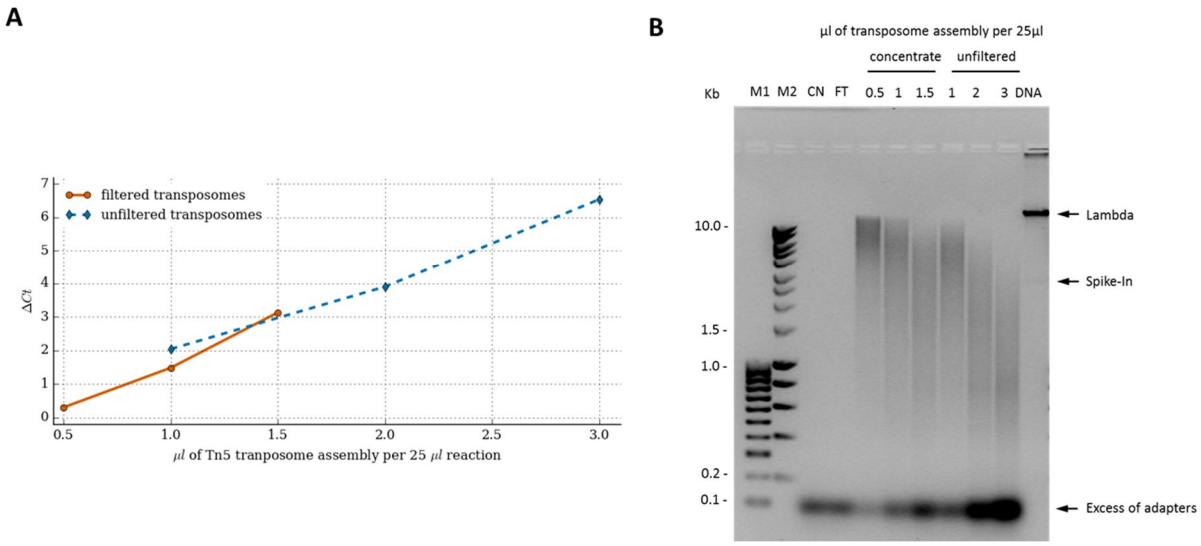


Figure 23 Purification of in-house Tn5 transposome assemblies with Amicon 100K ultra-0.5 filter units. 1.25μM Tn5ME-A transposome assembly was prepared by incubating Tn5 transposase and Tn5ME-A adapter for 30 min at 25°C. 20μl aliquot of transposomes was 1:1 diluted with 1X Exchange Buffer (50% glycerol) and 40μl diluted transposomes were loaded on filter unit and centrifuged at 5000g for 3min at 4°C. The concentrate volume was adjusted to 40μl with 1X Exchange Buffer (50% glycerol). 50ng of Lambda DNA was tagmented using a reference tagmentation template (pUC19/EcoRI) for 8min at 55°C with filtered and unfiltered transposomes (double concentrated). The reactions were stopped by adding 2% SDS to final concentration of 0.16% and heating for 7 mins at 55°C. The unpurified reaction products were analyzed by (A) FEA or (B) loading on 1.2% agarose gel. Markers: HyperLadder 100bp (M1), HyperLadder 1Kb (M2). Samples: filter concentrate (FC), filter flowthough (FT).

It is worth mentioning that a new approach for purification of Tn5 transposomes has been described [United States Patent 13/960,837]. The approach suggests the cleaning up of transposase complexes from a crude lysate, while they are immobilized on a solid support

through adapters. The author claims that his method has advantages over conventional purification procedures as transposase complex formation is carried out in more physiological conditions. It might be interesting to apply FEA for evaluation of this strategy.

In PITS procedure we performed tagmentation for 8 minutes at 55°C. Our test with a different tagmentation time showed that increase in this parameter can improve fragmentation efficiency. Incubating a reaction for 40 minutes gives DCt difference of about 2 (Figure 24). In comparison with other parameters, effect of tagmentation time is less expressed.

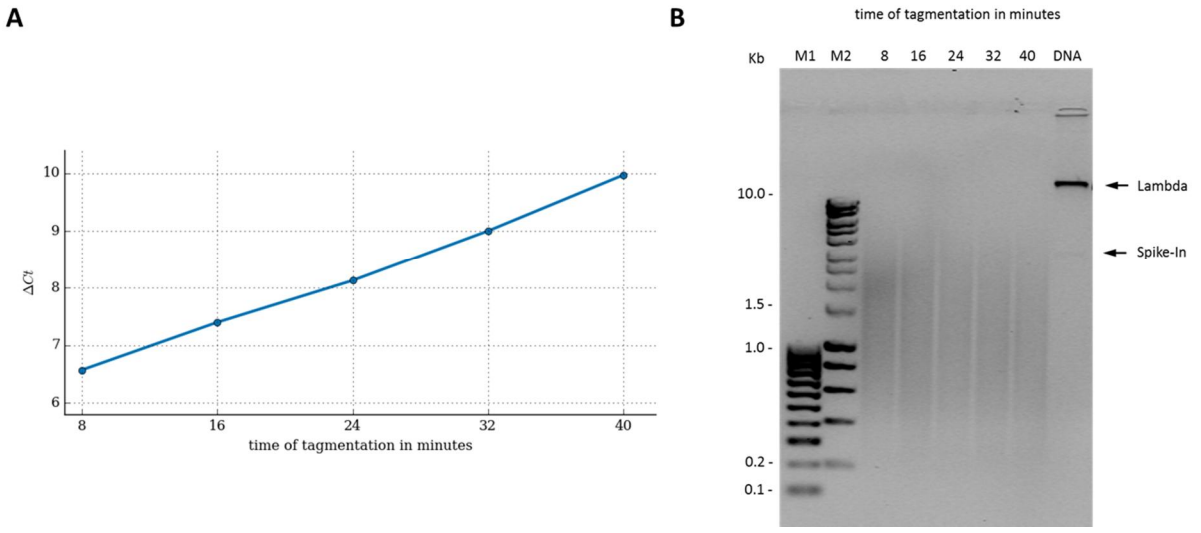


Figure 24 Evaluation of Tn5 transposome fragmentation efficiency by varying time of tagmentation reaction. 1.25µM Tn5 transposome assembly was prepared by incubating Tn5 enzyme with Tn5ME-A adapters for 30mins at 25°C. 50ng of Lambda DNA was tagmented along with a reference tagmentation template (pUC19/EcoRI) by adding 2.5µl of Tsomes with 8min divisible time intervals at 55°C. The tagmentation reactions were stopped by adding 2% SDS to final concentration of 0.16% and heating for 7mins at 55°C. The resulting mixtures were purified with AMPure DNA beads and analyzed: by (A) FEA or (B) on 1.1% agarose gel. Markers: HyperLadder 1kb (M1), HyperLadder 100bp (M2). Sample: tagmentation reaction without transposomes (DNA).

Earlier in *Rykalina et al.* we established that transposome assembly formation over time reaches a peak of maximum efficiency at 21 hours of incubation when monitoring during 24 hours. For PITS we prepare the transposome assembly incubation mixture for 21 hour at 25°C. As this finding was discovered for Tn5ME-A/B Tn5 transposomes, we carried out an experiment titrating transposomes assembled with Tn5ME-A type adapter. The transposomes proved to be active and the fragmentation efficiency was increasing under larger volumes (Figure 25).

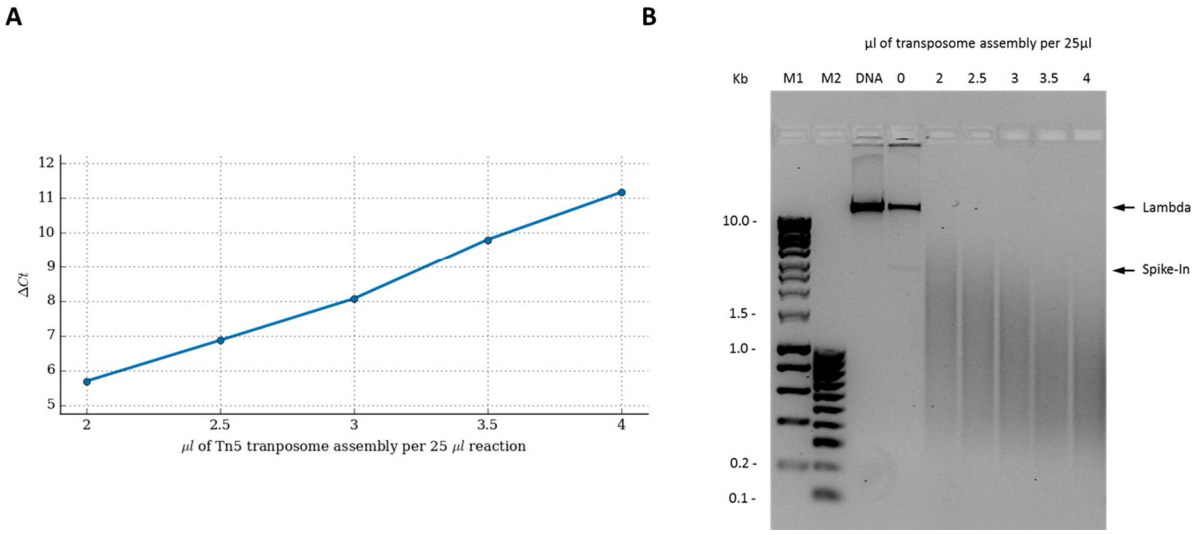


Figure 25 Evaluation of Tn5 transposome fragmentation efficiency by their volume in tagmentation reaction. 1.25μM Tn5 transposome assembly was prepared by incubating Tn5 enzyme with Tn5ME-A adapters for 21h at 25°C. 50ng of Lambda DNA was tagmented with a reference tagmentation template (pUC19/EcoRI) for 8min at 55°C. The tagmentation reactions were stopped by adding 2% SDS to final concentration of 0.16% and heating for 7mins at 55°C. The resulting mixtures were purified with AMPure DNA beads and analyzed: by (A) FEA or (B) loading on 1.1% agarose gel. Markers: Markers: HyperLadder 1kb (M1), HyperLadder 100bp (M2).

Reactions with working ranges of transposomes are shown in the Supplementary Figure 3.

The FEA technique is also applicable for evaluation of the stability of transposomes in time (Supplementary Table 3). Our tests confirmed the idea already mentioned in several papers that assembled Tn5 transposomes lose their activity during storage [Goryshin et al., 2000; Wang et al., 2013; Picelli et al., 2014].

Finally, Tn5 transposomes characterized by FEA could be used to prepare ATAC-seq libraries (Supplementary Methods Table 1, Figure 1). The results obtained by our collaborators, who were provided with in-house transposomes, showed that Nextera transposomes could be replaced by home-made assemblies. To obtain the same library quality as with a commercial kit the ATAC-seq protocol is still needed to be adapted. Some impediments are probably associated with an excess of unbound Tn5 adapters in in-house transposome batches. Probably, freezing the transposomes in liquid nitrogen (used for shipping) can influence their activity and this parameter should be under control.

Discussion of the PITS protocol

The topic of the PhD thesis relates to a novel (patented but not yet published) strategy for contiguity-preserving sequencing – paired indexing of fragmentation sites. A scheme, involving fragmentation with Tn5 transposase, was developed and tested – paired indexing of tagmentation sites (PITS). Properties of tagmentation reaction fit the requirements of the paired indexing procedure very well: the transposase inserts breaks into DNA and ligates adapters to the arising fragments before these fragments are physically separated. Asymmetrical cutting of the dsDNA strands by the transposase leaves complementary regions in the neighboring tagmentation fragments. Saving these regions during the library preparation process allows for their use as natural paired indices. We took advantage of this opportunity in our test experiments. Genomic DNA of phage λ was chosen as a test DNA material. We expected that, in the case of a high efficiency of the sequencing library preparation, it would be possible to reconstruct individual phage λ genomes taken for PITS, and in the case of a low efficiency – to assemble the phage sequence *de novo* using tagmentation fragments originating from different genomic copies, including those with the same genomic position.

Test experiments were performed using comparatively large number of genomic copies of phage λ. For a large percent of the fragments, a paired fragment could be found. When the scaffolds were assembled from the subset of the fragments with just one paired fragment, then, as expected, the smaller number of the genome copies was taken for library preparation. In this case the obtained scaffolds were longer and higher in number. So far, the longest scaffold we obtained consists of four unique fragments. To increase this number, we are going to further decrease the starting amount of genome copies. We are currently working on the development of an algorithm for assembling scaffolds from all fragments, including cases when more than two fragments share the same index. This should considerably increase the number and length of scaffolds.

Further work on the PITS protocol is dictated by the current bottleneck: incomplete sequencing of the pool of tagmentation fragments. Tagmentation gives a wide fragment size distribution, some out of sequencable range. Fragments get also lost because of an incomplete gap-repair, bias of amplification prior to sequencing, and bridge amplification on a flowcell surface. However, there are several possibilities to improve the protocol's performance. For example, chromatin can be trialled as an alternative tagmentation starting material. Transposase would cut between nucleosomes – this might reduce the number of fragments

below ~150nt (the length of the sequence protected by nucleosome) and also of excessively long fragments in the case when nucleosomes are positioned regularly. Amplification bias might be reduced if amplification is performed in emulsion. To improve the efficiency of the gap repair, it is possible to switch to transposon end Y- shape adapters containing both 1st and 2nd sequencing read primer regions. This would eliminate the need of oligonucleotide replacement step in the PTULI protocol.

It would also be advantageous for the protocol to preserve contiguity of the original DNA molecules not only for the dilution step, but until amplification. Using full transposon adapters with an uninterrupted sequence between the two ME regions would convert a DNA molecule into a DNA molecule with transposon inserts. Such a molecule can be processed through enzymatic reactions and purification steps without fragment losses. If there are fragments lost, then the whole molecule will be lost. We are currently working on transposition with full length transposons using a number of PCR products of different length, flanked with ME region for simulation. Though we have confirmed in a gel shift analysis that transposomes are formed, transposon insertion so far does not work properly (Supplementary Figures 5 and Supplementary Figure 6). We have also designed a full transposon structure allowing to insert paired indices together with the technical sequencing regions. We called such transposon a lampion because of the two bubble structures formed by non-complementary, single-stranded regions (Supplementary Figure 7A). Lampions allow attaching correct sequences to free 5´ and 3´ ends, arising at tagmentation sites, independently of the transposon orientation. Supplementary Figure 7 shows the scheme of the lampion preparation from oligonucleotides. We have prepared a transposon with the lampion structure (Supplementary Figure 8) and are going to test it for tagmentation as soon as tagmentation conditions are optimized using double-stranded transposons (PCR products).

The work on PITS proved to be very versatile. Apart from trialling the PITS itself, a lot of supporting work had to be done. A minor result is the preparation of the in-house Tn5 transposase. In-house enzyme is important not only for the further work on the PITS approach, but also for the currently running projects in our and neighbor laboratories, and for the establishment of new collaborations.

During the project two ready to use protocols were established – the PTULI library preparation method and the qPCR-based fragmentation efficiency assay (FEA) of Tn5 transposomes.

The PTULI sequencing library preparation protocol allows preparation of sequencing libraries from amol amounts of tagmentation products. We developed the accompanying procedure for

the subsequent non-residual loading of the library on the sequencer and detailed instructions on the control setup. Though developed for PITS approach, the PTULI library preparation might be used for other applications dependent on low amounts of starting material. We successfully used PTULI protocol for preparation of several test libraries, two of which were sequenced. To publish the protocol, it is still necessary to accurately determine its working range and trial it on genomic DNA other than that of phage λ.

The suggested qPCR-based fragmentation efficiency assay (FEA) is performed on a standard tagmentation template and uses fixed PCR detection region. It provides a reproducible system for the relative comparison of tagmentation reactions. We use this assay for titration of Tn5 batches and for optimization reactions. In principle, this approach of detecting fragmentation sites within a certain region may be used for assessment of other strategies, where the number of these sites is characteristic of the reaction. This includes non-transposase based fragmentation strategies, both enzymatic and physical, and also other reaction types, such as ligation. For site-specific reactions, e.g. restriction, our approach would even enable the use of absolute quantification of fragmentation sites to accurately measure the activity units as the percent of cut molecules in certain reaction conditions.

REFERENCES

Adams, D. S. (2003) Lab math: a handbook of measurements, calculations, and other quantitative skills for use at the bench. NY Cold Spring Harbor Laboratory Press.

Adey, A., Kitzman, J. O., Burton, J.N., Daza, R., Kumar, A., Christiansen, L., Ronaghi, M., Amini, S., Gunderson, K. L., Steemers, F. J. and Shendure, J. (2014) In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res* 24 (12), 2041-2049.

Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X. and Shendure, J. (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* 11 (12), R119.

Alkan, C., Sajjadian, S. and Eichler, E. E. (2011) Limitations of next generation genome assembly. *Nat Methods* 8 (1): 61-62.

Amini, S., Pushkarev, D., Christiansen, L., Kostem, E., Royce, T., Turk, C., Pignatelli, N., Adey, A., Kitzman, J. O., Vijayan, K., Ronaghi, M., Shendure, J., Gunderson, K. L. and Steemers, F. J. (2014) Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nature Genet* 46, 1343–1349.

Antonacci, F., Dennis, M. Y., Huddleston, J., Sudmant, P. H., Steinberg, K. M., Rosenfeld, J. A., Miroballo, M., Graves, T. A., Vives, L., Malig, M., Denman, L., Raja, A., Stuart, A., Tang, J., Munson, B., Shaffer, L. G., Amemiya, C. T., Wilson, R. K., and Eichler, E. E. (2014) Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. *Nat Genet* 46 (12), 1293-1302.

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

Belyaev A. S. (2014) Immobilized transposase complexes for DNA fragmentation and tagging. United States Patent 13/960,837.

Berlin, K., Koren, S., Chin, C. S., Drake, J. P., Landolin, J. M. and Phillippy, A. M. (2015) Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* 33 (6), 623-630.

Boettger, L. M., Handsaker, R. E., Zody, M. C. and McCarroll S. A. (2012) Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat Genet* 44 (8), 881-885.

Bogdanoff, D., Jou, T. and Lee, B. (2015) Characterization of action and efficiency of Tn5 Transposase through comparative qPCR. *JEMI* 19, 1-6.

Borodina, T., Adjaye, J. and Sultan, M. (2011) A strand-specific library preparation protocol for RNA sequencing. *Methods Enzymol* 500, 79-98.

Buenrostro, J. D., Giresi, P. G., Zaba, L.C., Chang, H.Y. and Greenleaf, W. J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10 (12), 1213-1218.

Burgtorf, C., Kepper, P., Hoehe, M., Schmitt, C., Reinhardt, R., Lehrach, H. and Sauer, S. (2003) Clone-based systematic haplotyping (CSH): a procedure for physical haplotyping of whole genomes. *Genome Res* 13 (12), 2717-2724.

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O. and Shendure, J. (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* 31 (12), 1119-1125.

Caruccio, N. (2011) Preparation of next-generation sequencing libraries using Nextera technology: simultaneous DNA fragmentation and adaptor tagging by in vitro transposition. *Methods Mol Biol* 733, 241–255.

Chaisson, M. J. P., Wilson, R. K. and Eichler E.E. (2015) Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet* 16 (11), 627-40.

Chaisson, M. J., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J. and Eichler, E. E. (2015) Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 517 (7536), 608-611.

Christiansen, L., Amini, S., Zhang, F., Ronaghi, M., Gunderson, K. L. and Steemers, F. J. (2017) Contiguity-Preserving Transposition Sequencing (CPT-Seq) for Genome-Wide Haplotyping, Assembly, and Single-Cell ATAC-Seq. *Methods Mol Biol* 1551, 207-221.

Clark, A. J. (2004) The role of haplotypes in candidate gene studies. *Genet Epidemiol* 27, 321-333.

De Jesus-Hernandez, M., Mackenzie, I. R., Boeve, B. F., Boxer, A. L., Baker, M., Rutherford, N. J., Nicholson, A. M., Finch, N. A., Flynn, H., Adamson, J. et al (2011) Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* 72 (2), 245-256.

de Vree, P. J. P., de Wit, E., Yilmaz, M., van de Heijning, M., Klous, P., Verstegen, M. J. A. M., Wan, Y., Teunissen, H., Krijger, P. H. L., Geeven, G. et al. (2014) Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nat Biotechnol* 32, 1019–1025.

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., Driscoll, M., Song, W., Kingsmore, S. F., Egholm, M. and Lasken R. S. (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* 99 (8), 5261-5266.

Dear, P. H. and Cook, P. R. (1989) Happy mapping: a proposal for linkage mapping the human genome. *Nucleic Acids Res* 17 (17), 6796-6807.

Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) Capturing chromosome conformation. *Science* 295 (5558), 1306-1311.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D. and Dekker, J. (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 16 (10), 1299-1309.

Douglas, J. A., Boehnke, M., Gillanders, E., Trent, J. M. and Gruber, S, B. (2001) Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nature Genet* 28, 361-364.

Drysdale, C. M., McGraw, D. W., Stack, C. B, Stephens, J. C., Judson, R. S., Nandabalan, K., Arnold, K., Ruano, G. and Liggett, S. B. (2000) Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc. Natl. Acad. Sci. USA* 97 (19), 10483–10488.

Eichler, E. E. (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet* 17 (11), 661-669.

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

Gloeckner, C., Kia, A., Bomati, E., He, M., Li, H., Kuersten, S., Osothprarop, T. F., Haskins, D., Burgess, J., Khanna, A., Schlingman, D., Vaidyanathan, R. (2015) Modified transposases for improved insertion sequence bias and increased DNA input tolerance. United States Patent US 2015/0291942 A1.

Glusman, G., Cox, H. C. and Roach, J. (2014) Whole-genome haplotyping approaches and genomic medicine. *Genome Med* 6 (9), 73.

Goryshin, I. Y., Jendrisak, J., Hoffman, L. M., Meis, R. and Reznikoff W. S. (2000) Insertional transposon mutagenesis by electroporation of released Tn5 transposition complexes. *Nat Biotechnol* 18 (1), 97-100.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U. and Kircher, M. (2010) A draft sequence of the Neandertal genome. *Science* 328 (5979), 710-722.

Groenendijk, M., Cantor, R.M., de Bruin, T.W. & Dallinga-Thie, G.M. The apoAI-CIII-AIV gene cluster. *Atherosclerosis* 157, 1–11 (2001).

Haque, F., Li, J., Wu, H., Liang, X. and Guo, P. (2013) Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA. *Nano Today* 8, 56–74.

Hardenbol, P., Banér, J., Jain, M., Nilsson, M., Namsaraev, E. A., Karlin-Neumann, G. A., Fakhrai-Rad. H., Ronaghi, M., Willis, T. D., Landegren, U. and Davis, R. W. (2003) Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol* 21 (6), 673-678.

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, A., Van Nieuwerburgh, F., Salomon, D. R., and Ordoukhanian P. (2014) Library constraction for next-generation sequencing: Overviews and challenges. *Biotechniques* 56 (2), 61-67. Bronner, I. F., Quail, M. A., Turner, D. J. and Swerdlow H. (2009) Improved protocols for Illumina sequencing. *Curr Protoc Hum Genet* doi:10.1002/0471142905.hg1802s62.

Heather, J. M. and Chain, B. (2016) The sequence of sequencers: The history of sequencing DNA. *Genomics* 107, 1–8.

Hou, Y., Fan, W., Yan, L., Li, R., Lian, Y., Huang, J., Li, J., Xu, L., Tang, F., Xie, X. S., and Qiao, J. (2013) Genome analyses of single human oocytes. *Cell* 155, 1492–1506.

Huang, X. (2015) Duplicating DNA with contiguity barcodes for genome and epigenome sequencing. International Patent PCT/US2014/065491.

Hurd, P. J. and Nelson, C. J. (2009) Advantages of next-generation sequencing versus the microarray in epigenetic research. *Brief Funct Genomic Proteomic* 8 (3), 174-183.

Jeffreys, A. J., Neumann, R. and Wilson, V. (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* 60 (3), 473-485.

Kaper, F., Swamy, S., Klotzle, B., Munchel, S., Cottrell, J., Bibikova, M., Chuang, H. Y., Kruglyak, S., Ronaghi, M., Eberle, M. A. and Fan, J. B. (2013) Whole-genome haplotyping by dilution, amplification, and sequencing. *Proc Natl Acad Sci U S A* 110 (14), 5552-5557.

Kaplan, N. and Dekker, J. (2013) High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat Biotechnol* 31 (12), 1143–1147.

Kia, A., Gloeckner, C., Osothprarop, T., Gormley, N., Bomati, E., Stephenson, M., Goryshin, I., He, M. M. (2017) Improved genome sequencing using an engineered transposase. *BMC Biotechnol* 17 (1): 6.

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. and Vogelstein, B. (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A* 108 (23), 9530-9535.

Kirby, A., Gnirke, A., Jaffe, D. B., Barešová, V., Pochet, N., Blumenstiel, B., Ye, C., Aird, D., Stevens, C., Robinson, J. T. et al. (2013) Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. *Nat Genet* 45 (3), 299-303.

Kirkness, E. F., Grindberg, R. V., Yee-Greenbaum, J., Marshall, C. R., Scherer, S. W., Lasken, R. S. and Venter, J. C. (2013) Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome Res* 23, 826–832.

Kitzman, J. O., Mackenzie, A. P., Adey, A., Hiatt, J. B., Patwardhan, R. P., Sudmant, P. H., Ng, S. B., Alkan, C., Qiu, R., Eichler, E. E. and Shendure J. (2011) Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol* 29 (1), 59-63.

Korbel, J. O. and Lee, C. (2013) Genome assembly and haplotypingwith Hi-C. *Nat Biotechnol* 31 (12), 1099-1101.

Kuhn, H. and Frank-Kamenetskii, M.D. (2005) Template-independent ligation of single-stranded DNA by T4 DNA ligase. *FEBS J* 272 (23), 5991-6000.

Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., Kertesz, M. and Snyder, M. (2014) Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol* 32 (3), 261-266.

Lam, E. T., Hastie, A., Lin., C., Ehrlich, D., Das, S. K., Austin, M. D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M. et al. (2012) Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat Biotechnol* 30 (8), 771-776.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh W. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409 (6822), 860-921.

Lawson, D. J., Hellenthal, G., Myers, S. and Falush, D. (2012) Inference of population structure using dense haplotype data. *PLoS Genet* 8 (1), e1002453.

Lemmers, R. J., Tawil, R., Petek, L. M., Balog, J., Block, G. J., Santen, G. W., Amell, A. M., van der Vliet, P. J., Almomani, R., Straasheijm, K. R. et al. (2012) Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat Genet* 44 (12), 1370-1374.

Levy, S. E. and Myers, R. (2016) Advancements in next-generation sequencing. *Annu Rev Genom Hum Genet* 17, 95-115.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G. et al. (2007) The diploid genome sequence of an individual human. *PLoS Biol* 5 (10), e254.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. and Dekker, J. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326 (5950), 289-293.

Liu, N., Zhang, K. and Zhao, H. (2008) Haplotype-association analysis. *Adv Genet* 60, 335-405.

Lo, C., Liu, R., Lee, J., Robasky, K., Byrne, S., Lucchesi, C., Aach, J., Church, G., Bafna, V. and Zhang, K. (2013) On the design of clone-based haplotyping. *Genome Biol* 14 (9), R100.

Lu, S., Zong, C., Fan, W., Yang, M., Li, J., Chapman, A. R., Zhu, P., Hu, X., Xu, L., Yan, L., Bai, F., Qiao, J., Tang, F., Li, R. and Xie, X. S. (2012) Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* 338, 1627–1630.

Ma, L., Xiao, Y., Huang, H., Wang, Q., Rao, W., Feng, Y., Zhang, K. and Song, Q. (2010) Direct determination of molecular haplotypes by chromosome microdissection. *Nat Methods* 7, 299–301.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A. et al. (2009) Finding the missing heritability of complex deseases. *Nature* 461 (7265), 747-753.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z. S., Munro, H. M., Abecasis, G. R. and Donnelly, P.; International HapMap Consortium (2008) A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* 78 (3), 437-50.

Morey, M., Fernández-Marmiesse, A., Castiñeiras, D., Fraga, J. M., Couce, M. L. and Cocho, J.A. (2013) A glimpse into past, present, and future DNA sequencing. *Mol Genet Metab* 110 (1-2), 3-24.

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E. T., Hastie, A. R., Marks, P., Lee, J., Chu, C., Lin, C., Džakula, Ž., et al. (2016) A hybrid approach for *de novo* human genome sequence assembly and phasing. *Nat Methods* 13 (7), 587-590.

Nievergelt, C. M., Libiger, O. and Schork, N. J. (2007) Generalized analysis of molecular variance. *PLoS Genet* 3, e51.

Ortigao, J. F. R., Rosch, H., Selter, H., Frohlich, A., Lorenz, A., Montenarh M. and Seliger H. (1992) Antisense effect of oligodeoxynucleotides with inverted terminal internucleotidic linkages: a minimal modification protecting against nucleolytic degradation. *Antisense Res & Dev* 2, 129-146.

Pareek, C. S., Smoczynski R. and Tretyn A. (2011) Sequencing technologies and genome sequencing. *J Appl Genet* 52 (4), 413-435.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H. and Soldatov, A. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* 37 (18), e123.

Paul, P. and Apgar, J. (2005) Single-molecule dilution and multiple displacement amplification for molecular haplotyping. *Biotechniques* 38 (4) 553-554, 556, 558-559.

Pendleton, M., Sebra, R., Pang, A. W., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Anantharaman, T., Hastie, A. et al. (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* 12 (8), 780-786.

Peters, B. A., Kermani, B. G., Sparks, A. B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y. T., Haas, J., Robasky, K., Zaranek, A. W., Lee, J. H., Ball, M. P., Peterson, J. E., Perazich, H., Yeung, G., Liu, J., Chen, L., Kennemer, M. I., Pothuraju, K., Konvicka, K., Tsoupko-Sitnikov, M., Pant, K. P., Ebert, J. C., Nilsen, G. B., Baccash, J., Halpern, A. L., Church, G. M. and Drmanac, R. (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 487, 190–195.

Picelli, S., Björklund, A. K., Reinius, B., Sagasser, S., Winberg, G. and Sandberg, R. (2014) Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res* 24 (12), 2033-2040.

Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., Troll, C. J., Fields, A., Hartley, P. D., Sugnet, C. W., Haussler, D., Rokhsar, D. S. and Green, R. E. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* 26, 342–350.

Pyo, C. W., Wang, R., Vu, Q., Cereb, N., Yang, S.Y., Duh, F. M., Wolinsky, S., Martin, M. P., Carrington, M. and Geraghty, D. E. (2013) Recombinant structures expand and contract inter and intragenic diversification at the KIR locus. *BMC Genomics* 14, 89.

Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., Swerdlow, H. and Turner, D. J. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5 (12), 1005-1010.

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., Bore, J. A., Koundouno, R., Dudas, G., Mikhail, A. et al. (2016) Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530 (7589), 228-32.

Raymond, C. K., Kas, A., Paddock, M., Qiu, R., Zhou, Y., Subramanian, S., Chang, J., Palmieri, A., Haugen, E., Kaul, R. and Olson M. V. (2005) Ancient haplotypes of the HLA Class II region. *Genome Res* 15 (9), 1250-1257.

Renton, A. E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J. R., Schymick, J. C., Laaksovirta, H., van Swieten, J. C., Myllykangas, L. et al. (2011) A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* 72 (2), 257-268.

Reuter, J. A., Spacek, D. V. and Snyder, M. P. (2015) High-throughput sequencing technologies. *Mol Cell* 58 (4), 586-597.

Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013) The advantages of SMRT sequencing. *Genome Biol* 14 (7), 405.

Ruano, G., Kidd, K. K. and Stephens, J. C. (1990) Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. Proc *Natl Acad Sci U S A* 87 (16), 6296-3600.

Ruijter, J. M., Pfaffl, M. W., Zhao, S., Spiess, A. N., Boggy, G., Blom, J., Rutledge, R. G., Sisti, D., Lievens, A., De Preter, K., Derveaux, S., Hellemans, J. and Vandesompele, J. (2013) Evaluation of qPCR curve analysis methods for reliable biomarker discovery: bias, resolution, precision, and implications. *Methods* 59 (1), 32-46.

Ryan, D. P., Dias da Silva, M. R., Soong, T. W., Fontaine, B., Donaldson, M. R., Kung, A. W. C., Jongjaroenprasert, W., Liang, M. C., Khoo, D. H.C., Cheah, J. S., Ho, S. C., Bernstein, H. S., Macie, R. M. B., Brown, R. H. J. and Ptáčcek, L. J. (2010) Mutations in potassium channel Kir2.6 cause susceptibility to thyrotoxic hypokalemic periodic paralysis. *Cell* 140 (1), 88-89.

Rykalina, V. N., Shadrin, A. A., Amstislavskiy, V. S., Rogaev, E. I., Lehrach, H. and Borodina, T., A. (2014) Exome sequencing from nanogram amounts of starting DNA: comparing three approaches. PLoS One 9 (7): e101154.

Rykalina, V. N., Shadrin, A. A., Lehrach, H. and Borodina, T. A. (2017) qPCR-based characterization of DNA fragmentation efficiency of Tn5 transposomes. *Biol Method Protoc* 2 (1): bpx001.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A.,

Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y. and Hu, H. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918.

Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelsto, J., Pääbo, S., Patterson, N. and Reich, D. (2014) The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507, 354–357.

Schiffels, S. and Durbin, R. (2014) Inferring human population size and separation history from multiple genome sequences. *Nature Genet* 46, 919–925.

Schwartz, D. C., Li, X., Hernandez, L.I., Ramnarain, S. P., Huff, E. J. and Wang, Y. K. (1993) Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science* 262 (5130), 110-114.

Schwartz, J. J., Lee, C., Hiatt, J. B., Adey, A. and Shendure J. (2012) Capturing native long-range contiguity by in situ library construction and optical sequencing. *Proc Natl Acad Sci U S A* 109 (46), 18749–18754.

Seliger, H., Frohlich, A., Montenarh, M., Ortigao J. F. R. and Rosch, H. (1991) Oligonucleotide analogs with terminal 3'-3'-internucleotidic and 5'-5'-internucleotidic linkages as antisense inhibitors of viral gene-expression**.** *Nucleosides & Nucleotides* 10, 469-477.

Selvaraj, S., R Dixon, J., Bansal, V. and Ren, B. (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* 31 (12), 1111–1118.

Seo, J. S., Rhie, A., Kim, J., Lee, S., Sohn, M. H., Kim, C. U., Hastie, A., Cao, H., Yun, J. Y., Kim, J., Kuk, J., Park, G. H., Kim, J., Ryu, H., Kim, J., Roh, M., Baek, J., Hunkapiller, M. W., Korlach, J., Shin, J. Y. and Kim, C. (2016) De novo assembly and phasing of a Korean human genome. *Nature* 538, 243-247.

Shaw, J. P., Kent, K., Bird, J., Fishback, J. and Froehler B. (1991) Modified deoxyoligonucleotides stable to exonuclease degradation in serum. *Nucleic Acid Res* 19 (4), 747-750.

Shendure, J. A., Schwarz, J. J., Aday, A. C., Lee, C. I., Hiatt, J. B., Kitzman, J. O. and Kumar, A. (2012) Massively parallel contiguity mapping. International Patent PCT/US2012/023679.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* 38 (11), 1348-1354.

Sohn, J. and Nam, J. (2016) The present and future of *de novo* whole-genome assembly. *Brief Bioinform* 1-18.

Sram, J., Sommer, S. S., Liu, Q. (2008) Microarray-based DNA re-sequencing using 3' blocked primers. *Anal Biochem* 3744, 41-47.

Steemers, F. J., Fisher, J. S., Gunderson, K. L., Amini, S. and Gloeckner, C. Methods and compositions using one-sided transposition. (2016) International Patent PCT/US2015/038050.

Steemers, F. J., Gunderson, K., Royce, T., Pignatelli, N., Goryshin, I. Y., Caruccio, N., Maffitt, M., Jendrisak, J., Amini, S., Kaper, F., Turk, C. and Kahlor, R. (2012) Linking sequence reads using paired code tags. International Patent PCT/US201 1/059642.

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J.; 1000 Genomes Project and Eichler, E. E. (2010) Diversity of human copy number variation and multicopy genes. *Science* 330 (6004), 641-646.

Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. and Schork, N. J. (2011) The importance of phase information for human genomics. *Nat Rev Genet* 12, 215–223.

The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. (2010) *Nature* 467 (7319), 1061-73.

The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1.092 human genomes. (2012) *Nature* 491 (7422), 56-65.

Valouev, A., Schwartz, D. C., Zhou, S. and Waterman, M. S. (2006) An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proc Natl Acad Sci U S A* 103 (43), 15770-15775.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., Thermes, C. (2014) Ten years of next-generation sequencing technology. *Trends Genet* 30 (9), 418-426.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A. et al. (2001) The sequence of the human genome. *Science* 291(5507), 304-1351.

Vernot, B. and Akey, J. M. (2014) Resurrecting surviving Neandertal lineages from modern human genomes. *Science* 343, 1017–1021.

Wang, J., Fan, H. C., Behr, B. and Quake, S. R. (2012) Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* 150, 402–412.

Wang, Q., Gu, L., Adey, A., Radlwimmer, B., Wang, W., Hovestadt, V., Bähr, M., Wolf, S., Shendure, J., Eils, R., Plass, C. and Weichenhan, D. (2013) Tagmentation-based whole-genome bisulfite sequencing. *Nat Protot* 8 (10), 2022-2032.

www.illumina.com

www.molbiol.ru

www.pacificbiosciences.com

Yan, H., Papadopoulos, N., Marra, G., Perrera, C., Jiricny, J., Boland, C. R., Lynch, H. T., Chadwick, R. B., de la Chapelle, A., Berg, K., Eshleman, J. R., Yuan, W., Markowitz, S., Laken, S. J., Lengauer, C., Kinzler, K. W. and Vogelstein, B. (2000) Conversion of diploidy to haploidy. *Nature* 403 (6771), 723-4.

Yang, H., Chen, X. and Wong, W. H. (2011) Completely phased genome sequencing through chromosome sorting. *Proc Natl Acad Sci USA* 108, 12–17.

Zhang, K., Zhu, J., Shendure, J., Porreca, G. J., Aach, J. D., Mitra, R. D. and Church, G. M. (2006) Long-range polony haplotyping of individual human chromosome molecules. *Nature Genet* 38, 382-387.

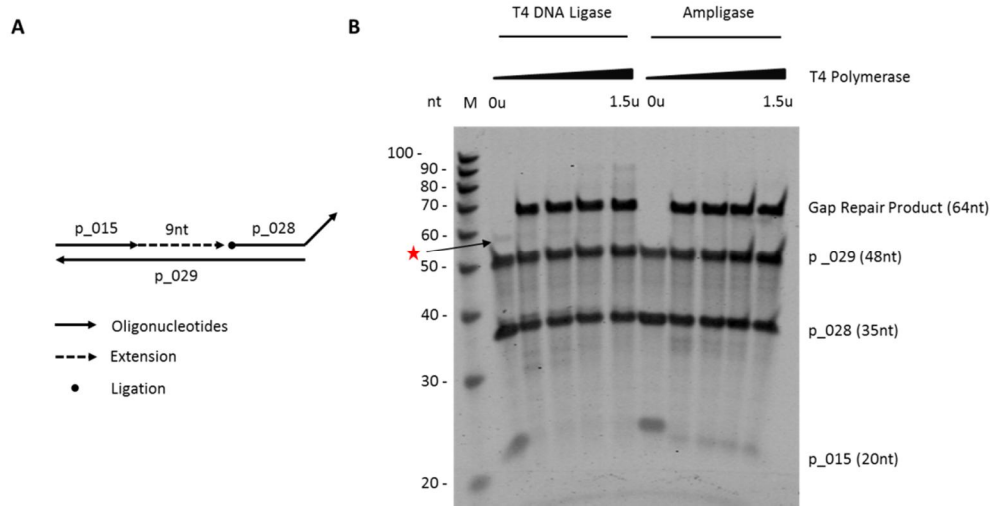
Zheng, G. X. Y., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M. et al. (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* 34 (3), 303-311.

Zody, M. C., Jiang, Z., Fung, H. C., Antonacci, F., Hillier, L. W., Cardone, M. F., Graves, T. A., Kidd, J. M., Cheng, Z., Abouelleil, A., Chen, L., Wallis, J., Glasscock, J., Wilson, R. K., Reily, A. D., Duckworth, J., Ventura, M., Hardy, J., Warren, W. C. and Eichler, E. E. (2008) Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet* 40 (9), 1076-1083.

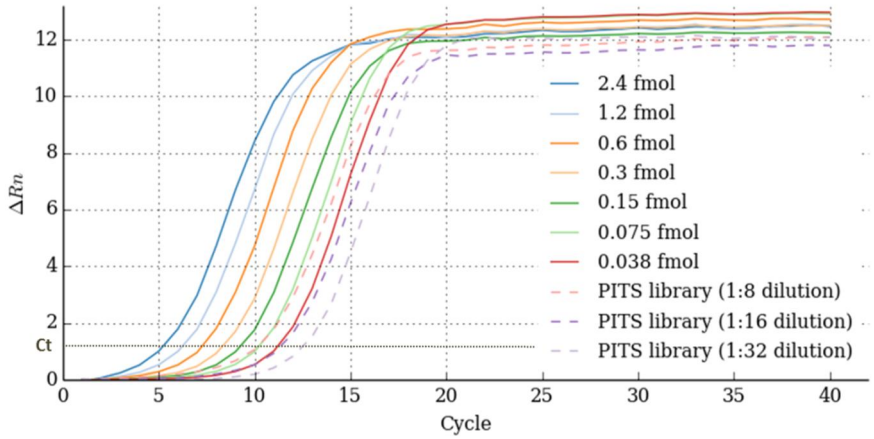
Zong, C., Lu, S., Chapman, A. R. and Xie, X. S. (2012) Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 338 (6114), 1622-1626.

SUPPLEMENTARY

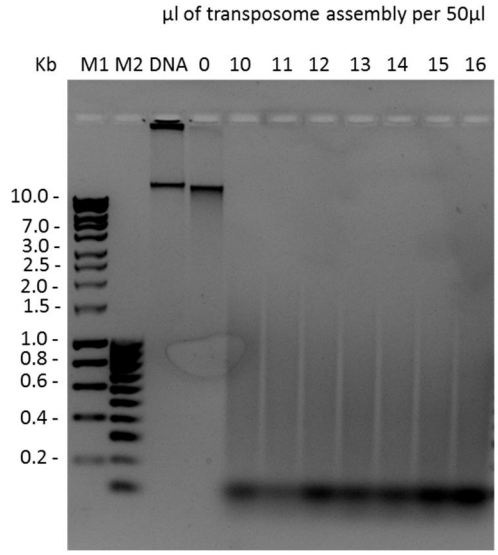
Supplementary Figures



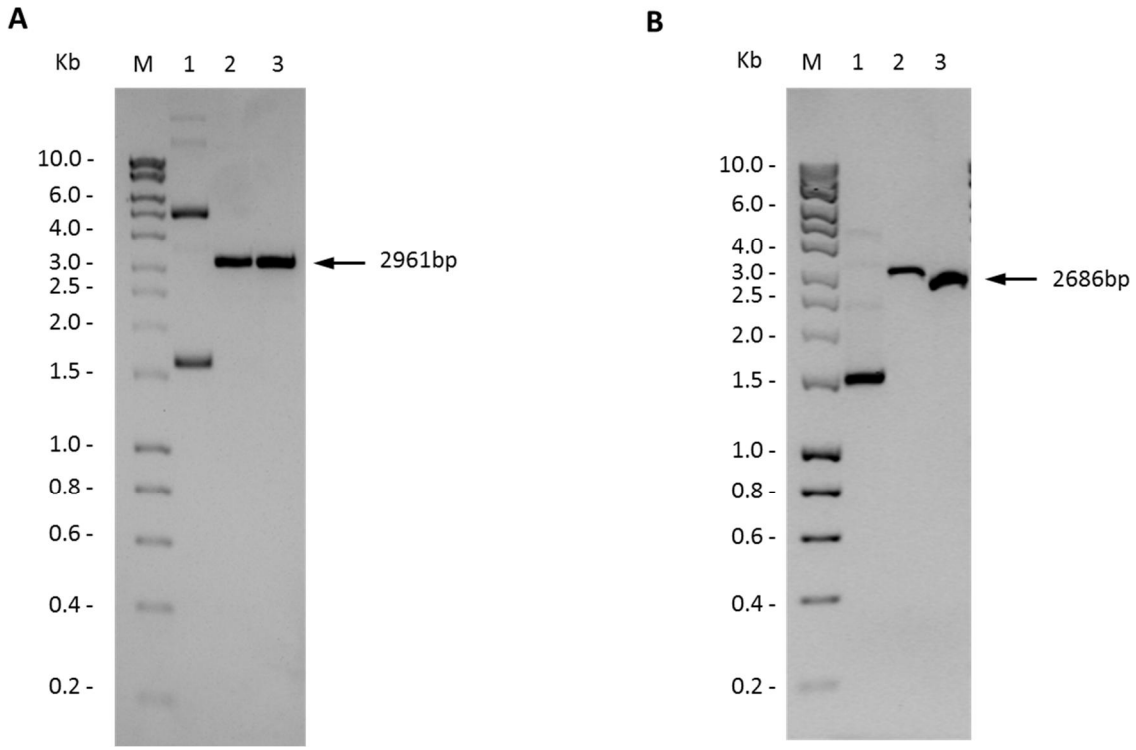
Supplementary Figure 1 (A) Gap repair oligonucleotide model used for reaction setup. Oligonucleotides p_015 and p_028 are annealed to p_029 forming a 9nt gap. (B) Comparison of T4 DNA Ligase and Ampligase performance in the gap repair. 20pmol of the p_015/p_028/p_029 complex, 400 units of T4 DNA Ligase or 100 units of Ampligase were taken per 10µl reaction; T4 DNA Polymerase was titrated by 1:2 series dilution. Gap repair was carried out for 30 min at 16°C (with T4 DNA Ligase) or at 37°C (with Ampligase). Half of the reaction was mixed with 2x Oligo Loading Buffer, heated at 95°C for 5 min and visualized on 10% PAGE (130V for 1.15h) along with the Ladder 20/100 (M). Red asterisk points to the non-template ligation product of p_015 and p_028.



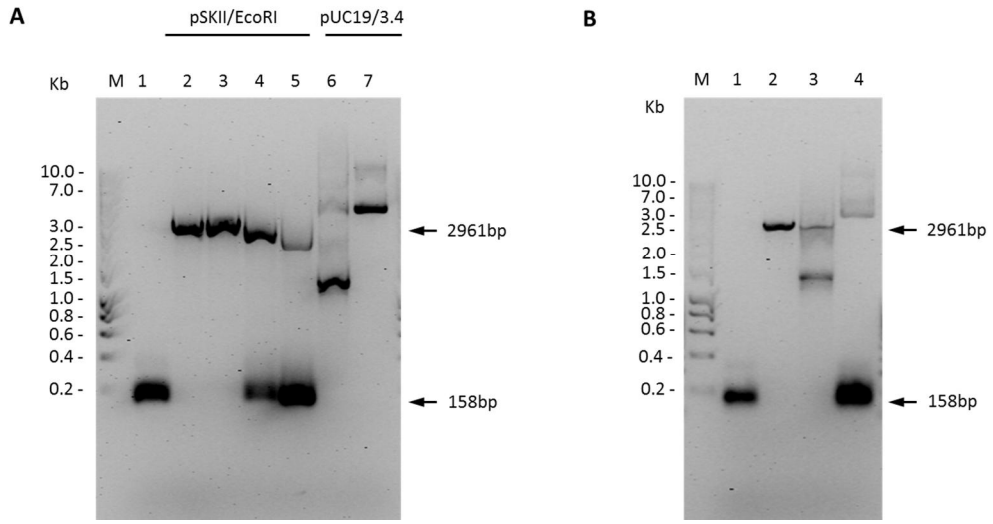
Supplementary Figure 2 Adjustment of amplification cycles amount for a PITS library according to the PTULI protocol. Test amplification library (here – PITS library) was amplified with 23 cycles, purified and brought to the desired 5µl volume. 1/8µl, 1/16µl and 1/32µl of the library were amplified along the 1:2 dilution series of the reference library (corresponding amount of starting material is indicated in the legend). The required concentration of the PITS library is 1.2fmol/µl. If the amplification was sufficient, the 1/8µl, 1/16µl and 1/32µl of the PITS library would have raised as 0.15, 0.075 and 0.038 fmol respectively. According to the amplification plot, one more cycle is required for amplification of the PITS library. Indeed, the main sample was amplified with 24 cycles, and later performed as expected on the MiSeq flowcell (see the Results section for PITS 4500 library).



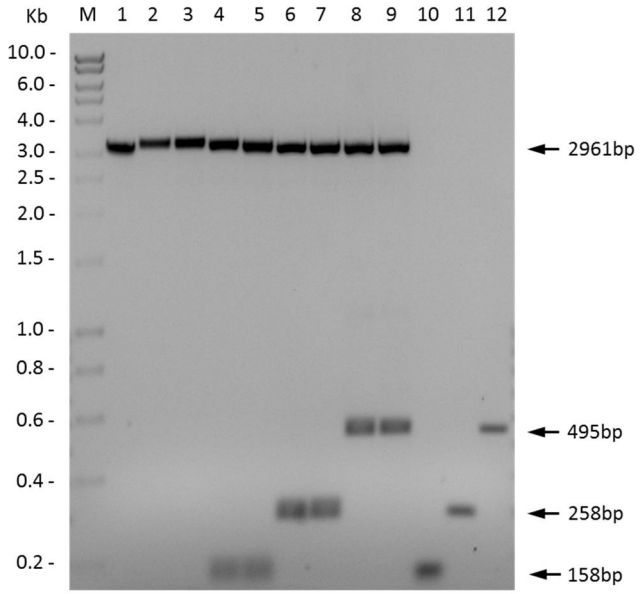
Supplementary Figure 3 Tagmentation of Lambda DNA varying the volume of transposome assembly in reaction (21h, 25°C). The tagmentation products (8min at 55°C) were purified with QIAquick PCR Amplification Kit and analyzed on 1.1% (100V for 2h). Markers: HyperLadder 1kb (M1), HyperLadder 100bp (M2).



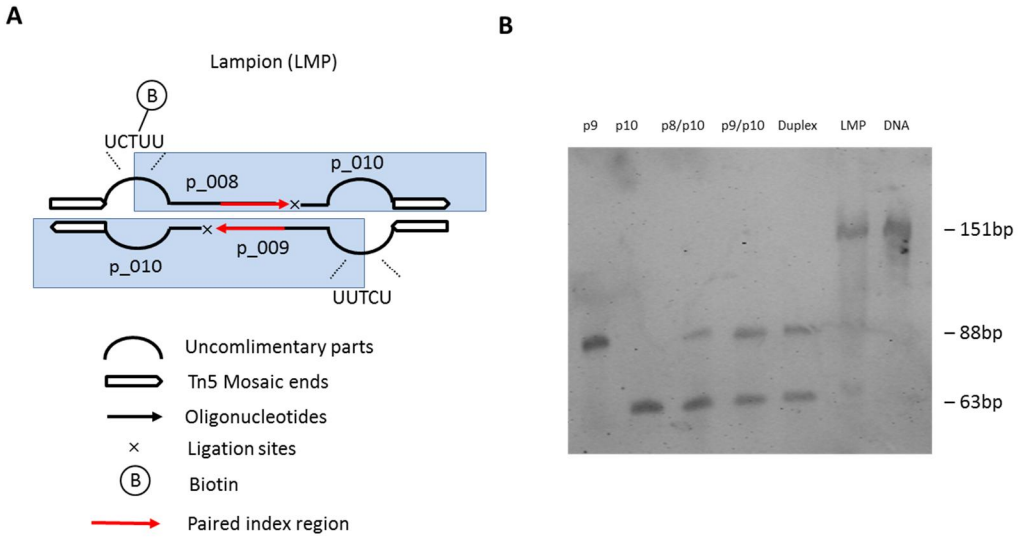
Supplementary Figure 4 Preparation of linearized plasmids. Sample containing 50ng of DNA analyzed with 1% agarose gel (120V for 2.5h). (A) Plasmid pKSII. Marker: HyperLadder 1kb (M), Lane 1 – pKSII (intact), Lane 2 – unpurified pKSII /EcoRI restriction mixture, Lane 3 – column purified pKSII /EcoRI. (B) Plasmid pUC19. Marker: HyperLadder 1kb (M), Lane 1 – pUC19, Lane 2 – unpurified pUC19/EcoRI, Lane 3 – column purified pUC19/EcoRI.



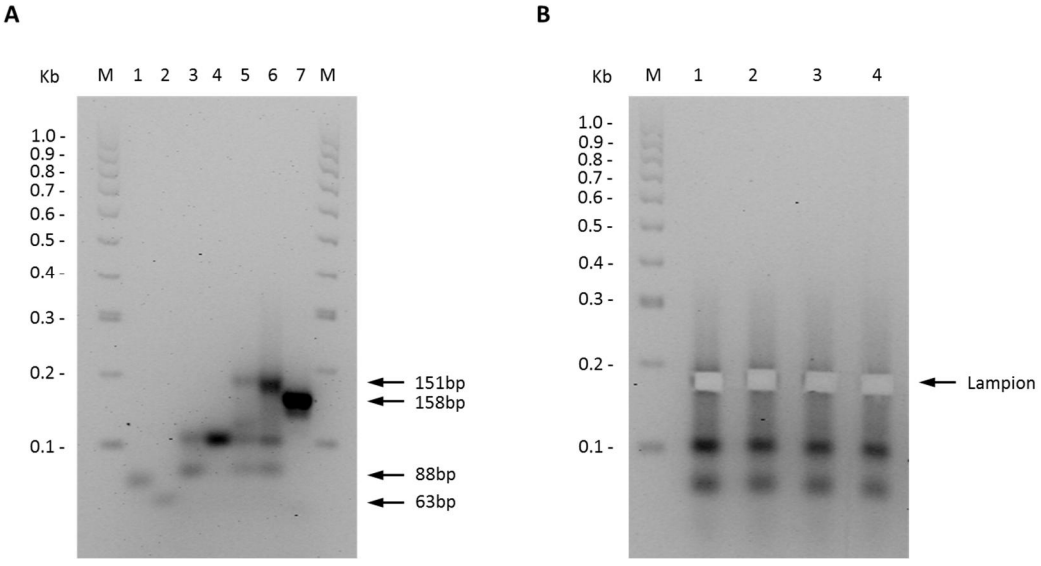
Supplementary Figure 5 Tagmentation of pKSII/EcoRI or pUC19/3.4 (from the Epicentre EZ-Tn5 Kit) with transposomes consisting of full transposon (158bp PCR product transposon or 1221bp EZ-Tn5 KAN-2 Transposon) and transposase from Epicentre EZ-Tn5 Kit. (**A**) Lanes 1-5 correspond to158bp transposon and pKSII/EcoRI tagmentation template: Lane 1 –158bp transposon , Lane 2 – pKSII/EcoRI, Lane 3 – transposomes (-) tagmentation control, Lane 4, 5 – tagmentation with 1:5 and 1:15 ratios of target DNA to transposon. Lanes 6 and 7 correspond to 1221bp EZ-Tn5 KAN-2 Transposon and pUC19/3.4 tagmentation template: Lane 6 – transposomes (-) tagmentation control, Lane 7 – tagmentation with 1:5 ratio of target DNA to transposon. (B) Lane 1 – 158bp transposon after end-repair, Lane 2 – pKSII/EcoRI, Lane 3 – insertion of 1221bp EZ-Tn5 KAN-2 Transposon in pKSII/EcoRI (1:5 ratio of target DNA to transposon), Lane 4 – insertion of 158bp end-repaired transposon in pUC19/3.4 (1:5 ratio of target DNA to transposon). Preliminary results are unclear. Commercial system seems to work: though we do not observe a shift of the tagmentation template size due to insertion of transposons, all transposon seem to be inserted, because no transposon band is seen on the gel - A, Lane 7. However, substituting either of commercial reagents leads to a different result: when EZ-Tn5 KAN-2 Transposon is inserted in the pKSII/EcoRI, or 158bp PCR product transposon is inserted into the template from the kit, still a lot of transposon remains unused.



Supplementary Figure 6 Tagmentation of pKSII/EcoRI with transposomes consisting of full transposons of different lengths (158bp, 258bp, 495bp) and in-house Tn5 transposase. Lane 1- pKSII/EcoRI, Lane 2 and Lane 3 – transposomes (-) tagmentation controls, Lane 4, 5 – insertion of 158bp transposon, Lane 6, 7 – insertion of 258bp transposon, Lane 8 and Lane 9 – insertions of 495bp transposon, Lane 10 – 158bp transposon, Lane 11 – 258bp transposon, Lane 12 – 495bp transposon. No shift of the tagmentation template band is observed, independent of the transposon length. In-house transposase and the same tagmentation template performed well in the tagmentation experiments with transposon end adapters.



Supplementary Figure 7 Lampion: full transposon with internal uncomplementary regions. (A) Scheme of the lampion preparation from 3 oligonucleotides – p_008, p_009, p_010. Duplexes p_008/p_010 and p_009/p_010 anneal to each other by protruding 3‘ends, and the nicks are closed with a ligase. Resulting structure with two uncomplementary regions (shown as not parallel lines) is a lampion. Biotin is used for purification of the full length products. Paired indices containing region is shown with red color. Black lines correspond to the technical sequences required for amplification and sequencing. After tagmentation and filling of 9nt gaps, lampions are fully integrated in the tagmentation template. To fragment DNA, it is treated with UDGase which cuts out uridines within uridine-containing sequences (UUTCU) of the lampion. After that the transposon strands parts shown within blue rectangle remain bound to the 5‘ end of the fragments, those out of blue rectangle – remain attached to the 3‘ends. (B) Analytical PAGE gel showing steps of lampion preparation. Control DNA is loaded for size control.



Supplementary Figure 8 Preparative lampion structure isolation from an agarose gel. (A) Synthesis of lampion-like Tn5 adapter structure. Reaction aliquots (5pmol) were loaded on 2% agarose gel (low melting point agarose, 100V for 4h). Marker – HyperLadder 100bp (M), Lane 1 – oligonucleotide p_009 (88bp), Lane 2 - phosphorylated oligonucleotide p_010 (63bp), Lane 3 – duplex p_008/p_010, Lane 4 – duplex p_009/p_010, Lane 5 - mix of two duplexes without T4 DNA Ligase, Lane 6 - ligation of two duplexed, Lane 7 – double-stranded 158bp PCR product used as a size control. (B) Isolation from an agarose gel. Marker – HyperLadder 100bp (M), Lane 1 to Lane 4 are preparative samples for extraction.

Supplementary Tables

Supplementary Table 1 Index sequences in the sequencing reads.

Most represented 9nt index sequences		Most unrepresented 9nt index sequences	
Sequence	**Ratio**	**Sequence**	**Ratio**
GGGTAAAAC	36.19	GCGGCAATA	0.099
GTATAAAGC	34.99	TTTTGCTGG	0.099
GTGTATTAC	34.69	ACGGGAAAA	0.099
GAATTACAC	34.09	CTGGATGCA	0.099
GTCTTAAAC	34.09	ATGTCGATA	0.099
GACGTACAC	31.10	TTTGTCTTC	0.099
GTACTGGAG	30.20	TGTGGTGAT	0.074
CCTTTGTAC	29.91	CACCGCCAG	0.074
GTTCAGAGC	29.91	GCTGTCGCG	0.074
GAAGAAGCC	29.91	TGAACTGAT	0.074
GTGTTAATC	29.31	AATGGTTTC	0.074
GTGTGGGGG	29.31	GGGACGAAA	0.074
ATCTAACAC	29.31	TGCCAGCGA	0.074
GTATCAGTC	28.11	TGCCGGACA	0.074
CCTCTAAAT	27.51	CGGCGCGTT	0.074
GAATAACCA	27.21	CTGGCTGCA	0.074
CGGCTACTC	26.91	AACCGCTTC	0.074
GTGTGCTCC	26.62	GGCGCTGTA	0.074
CCTTGATAC	26.02	GCCGGACAG	0.074
GAGTACGGC	26.02	ACGCCCGGC	0.074
GTGTTGATC	25.42	GCAGGCAGA	0.074
GCTCCAGCC	25.42	GCAGGCAGA	0.059
GGAGAAATC	25.12	TCAGCCAGC	0.059
GTGTTATTC	24.82	GCTGACGTT	0.049
GGCGATACC	24.82	TTTTTTATA	0.049

The 25 most represented (left column) and 25 most underrepresented (right column) 9nt indices sequences in the sequencing reads. Numbers show ratio of the amount of a 9nt stretch in all non-duplicated reads from both libraries to the number of that 9nt stretch in the phage genome. Theoretically, if transposase acts randomly, this ratio should be close to 1 for all 9nt stretches. On practice, the situation is different.

Supplementary Table 2 Sequences of Tn5 transposase clones (C-1 and C-2), determined by Sanger sequencing.

Tn5 C-1 F_T7 (19..909)	CTCTAGAATAATTTTGTTAACTTTAAGGAAGGAGATATACATATGATTACCAGTGCACTGCATCGTGCGGC GGATTGGGCGAAAAGCGTGTTTTCTAGTGCTGCGCTGGGTGATCCGCGTCGTACCGCGCGTCTGGTGAATG TTGCGGCGCAACTGGCCAAATATAGCGGCAAAAGCATTACCATTAGCAGCGAAGGCAGCAAAGCCATGCAG GAAGGCGCGTATCGTTTTATTCGTAATCCGAACGTGAGCGCGGAAGCGATTCGTAAAGCGGGTGCCATGCA GACCGTGAAACTGGCCCAGGAATTTCCGGAACTGCTGGCAATTGAAGATACCACCTCTCTGAGCTATCGTC ATCAGGTGGCGGAAGAACTGGGCAAACTGGGTAGCATTCAGGATAAAAGCCGTGGTTGGTGGGTGCATAGC GTGCTGCTGCTGGAAGCGACCACCTTTCGTACCGTGGGCCTGCTGCATCAAGAATGGTGGATGCGTCCGGA TGATCCGGCGGATGCGGATGAAAAAGAAAGCGGCAAATGGCTGGCCGCTGCTGCAACTTCGCGTCTGAGAA TGGGCAGCATGATGAGCAACGTGATTGCGGTGTGCGATCGTGAAGCGGATATTCATGCGTATCTGCAAGAT AAACTGGCCCATAACGAACGTTTTGTGGTGCGTAGCAAACATCCGCGTAAAGATGTGGAAAGCGGCCTGTA TCTGTATGATCACCTGAAAAACCAGCCGGAACTGGGCGGCTATCAGATTAGCATTCCGCAGAAAGGCGTGG TGGATAAACGTGGCAAACGTAAAAACCGTCCGGCGCGTAAAGCGAGCCTGAACCTGCGTAGCGGCCGTATT ACCCTGAAACAGGGCAACATTACCCTGAACGCGGTGCTG
Tn5 C-1 R_premix (20..901)	TACTGTTGGGCCGCGCACCCGGCACGATGTCGGCGATGCGTACCGACTCGCCCTCGGGTAGGGCAACTAGT GCATCTCCCGTGATGCAGATTTTAATGCCCTGCGCCATCAGGTCTTTCGCGGCCAGAAAGCCATCCAGTTT GCTTTGCAGCGCTTCCCAACCTTCCCACAGCGCACCCCAGCTCGCAATGCCGGTACGTTTGCTATCCATAA AGCCGCCCAGACGCGCAATCGCCATATACGCCCATTGCAGGCTGCCCGCTTTTTCTTTGCGTTTGCGTTTG CCTTTATCCAGATAGCCCAGCAGTTGGCATTCATCCGGGGTCAGCACGGTTTCCGCGCTCTGGCTTTCAAC GTGTTCCGCTTCTTTCAGCAGGCCCTGCGCACGCAGTGCTTGCGGCGGAGTAAAAGATTCACGCAGTTGCA GCAGACGCACCGCCACAAAGCTCAGAATGCTCACCATACGTTCCAGGTTATCCGGTTCTTCCATACGCTGA CGTTCCGCACCCGCACCCGTTTTCCACGCTTTGTGAAATTCTTCAATGCGCCAACGATGGGTATAAATATC AATCACACGCAGCGCTTGGGCCAGACTTTCCACCGGCTCGCTGGTCAGCAGCAGCCATTTCAGCGGGGTTT CGCCTTTCGGCGGATTAATTTCTTCGGCCAGCACCGCGTTCAGGGTAATGTTGCCCTGTTTCAGGGTAATA CGGCCGCTACGCAGGCTCAGGCTCGCTTTACGCGCCGGACGGTTTTTACGTTTGCCACGTTTATCCACCAC GCCTTTCTGCGGAATGCTAATCTGATAGCCGCCCAGTTCCGGCTGGTTTTTCAGGTGATCATACAGATACA GGCCGCTTTCCACATCTTTACGCGGATGTT
Tn5 C-2 F_T7 (23..923)	TCTAGAATAATTTTGTTAACTTTAAGAAGGAGATATACATATGATTACCAGTGCACTGCATCGTGCGGCGG ATTGGGCGAAAAGCGTGTTTTCTAGTGCTGCGCTGGGTGATCCGCGTCGTACCGCGCGTCTGGTGAATGTT GCGGCGCAACTGGCCAAATATAGCGGCAAAAGCATTACCATTAGCAGCGAAGGCAGCAAAGCCATGCAGGA AGGCGCGTATCGTTTTATTCGTAATCCGAACGTGAGCGCGGAAGCGATTCGTAAAGCGGGTGCCATGCAGA CCGTGAAACTGGCCCAGGAATTTCCGGAACTGCTGGCAATTGAAGATACCACCTCTCTGAGCTATCGTCAT CAGGTGGCGGAAGAACTGGGCAAACTGGGTAGCATTCAGGATAAAAGCCGTGGTTGGTGGGTGCATAGCGT GCTGCTGCTGGAAGCGACCACCTTTCGTACCGTGGGCCTGCTGCATCAAGAATGGTGGATGCGTCCGGATG ATCCGGCGGATGCGGATGAAAAAGAAAGCGGCAAATGGCTGGCCGCTGCTGCAACTTCGCGTCTGAGAATG GGCAGCATGATGAGCAACGTGATTGCGGTGTGCGATCGTGAAGCGGATATTCATGCGTATCTGCAAGATAA ACTGGCCCATAACGAACGTTTTGTGGTGCGTAGCAAACATCCGCGTAAAGATGTGGAAAGCGGCCTGTATC TGTATGATCACCTGAAAAACCAGCCGGAACTGGGCGGCTATCAGATTAGCATTCCGCAGAAAGGCGTGGTG GATAAACGTGGCAAACGTAAAAACCGTCCGGCGCGTAAAGCGAGCCTGAACCTGCGTAGCGGCCGTATTAC CCTGAAACAGGGCAACATTACCCTGAACGCGGTGCTGGCCGAAAAAATT
Tn5 C-2 R_premix (18..897)	GTATGTTGGGCCGCGCACCCGGCACGATGTCGGCGATGCGTACCGACTCGCCCTCGGGTAGGGCAACTAGT GCATCTCCCGTGATGCAGATTTTAATGCCCTGCGCCATCAGGTCTTTCGCGGCCAGAAAGCCATCCAGTTT GCTTTGCAGCGCTTCCCAACCTTCCCACAGCGCACCCCAGCTCGCAATGCCGGTACGTTTGCTATCCATAA AGCCGCCCAGACGCGCAATCGCCATATACGCCCATTGCAGGCTGCCCGCTTTTTCTTTGCGTTTGCGTTTG CCTTTATCCAGATAGCCCAGCAGTTGGCATTCATCCGGGGTCAGCACGGTTTCCGCGCTCTGGCTTTCAAC GTGTTCCGCTTCTTTCAGCAGGCCCTGCGCACGCAGTGCTTGCGGCGGAGTAAAAGATTCACGCAGTTGCA GCAGACGCACCGCCACAAAGCTCAGAATGCTCACCATACGTTCCAGGTTATCCGGTTCTTCCATACGCTGA CGTTCCGCACCCGCACCCGTTTTCCACGCTTTGTGAAATTCTTCAATGCGCCAACGATGGGTATAAATATC AATCACACGCAGCGCTTGGGCCAGACTTTCCACCGGCTCGCTGGTCAGCAGCAGCCATTTCAGCGGGGTTT CGCCTTTCGGCGGATTAATTTCTTCGGCCAGCACCGCGTTCAGGGTAATGTTGCCCTGTTTCAGGGTAATA CGGCCGCTACGCAGGCTCAGGCTCGCTTTACGCGCCGGACGGTTTTTACGTTTGCCACGTTTATCCACCAC GCCTTTCTGCGGAATGCTAATCTGATAGCCGCCCAGTTCCGGCTGGTTTTTCAGGTGATCATACAGATACA GGCCGCTTTCCACATCTTTACGCGGATG

Supplementary Table 3 Correlation of storage time with in-house transposome activity loss.

Time Intervals (days)	pUC19/EcoRI (prep 1) (ΔCt)	pUC19/EcoRI (prep 2) (ΔCt)
1	6.08	8.68
7	4.92	7.77
90	3.64	5.85

Tn5 transposome batches were prepared by mixing equal volumes of 20µM Tn5 enzyme and 20µM Tn5ME-A/B adapters and stored at -20°C during three time intervals. Tagmentation reactions were carried out by incubating the mixtures in the thermocycler for 30 min at 37°C. pUC19 plasmids (preparation 1 and 2) linearized with EcoRI was applied as a tagmentation template. Tagmentation of 50ng plasmid DNA was performed in 1X TB Buffer in total reaction volume of 50µl for 10 min at 58°C. 2µl of 2% SDS was added to each reaction, which was then incubated at 55°C for 7 min. Tagmentation reactions were purified with AMPure XP Beads according to manufacturer's instructions, with the following settings: beads were added to the Tn5 reactions at a 0.8:1 ratio; beads were washed with 70% freshly prepared ethanol; DNA was eluted in 50µl of EB Buffer. 1/300 aliquots of the purified tagmentation reactions were analyzed by qPCR-based FEA [Rykalina et al., 2017]. qPCR was conducted using the SYBR Green PCR Core Reagents and 1 unit of Immolase per reaction. Following temperature profile conditions were applied: 95°C for 10 min followed by 40 cycles of 95°C for 15 sec, 63°C for 15 sec and 72°C for 70 (1240bp PCR product). Each reaction contained 0.5µM forward (p_017) and 0.5µM reverse (p_018) primers, in a final reaction volume of 20µl. The storage conditions were tested for three time intervals: 1 day, 7 days and 90 days.

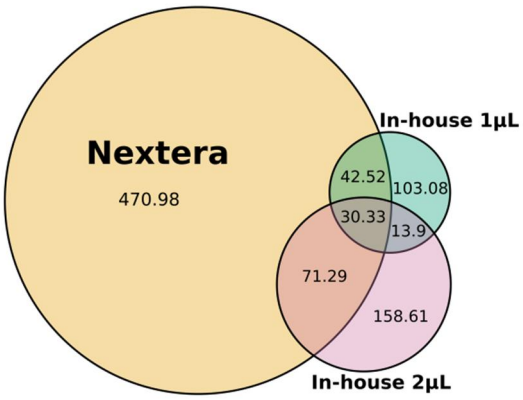
Supplementary Methods

Evaluation of in-house Tn5 Tsomes by ATAC-seq

Our colleagues were provided by us with a characterized aliquot of Tn5 In-house Transposomes. Transposomes were assembled with Illumina A/B adapters as described [Rykalina et al., 2017]. Activity of Tn5 In-house transposomes was assessed according to FEA protocol in comparison with TDE1 Tagment DNA Enzyme from Illumina Nextera DNA Library Preparation Kit. Transposome volume for tagmentation with In-house enzyme was adjusted using the following equation:

$$V_{\text{In-house}} = \frac{(1.4 \cdot V_{\text{Nextera}}) + 0.8}{2.1}$$

The In-house Transposomes were delivered frozen solid in liquid nitrogen. The influence of freezing conditions on Tn5 Transposome activity was not assessed. ATAC-seq assay was carried out for three samples on 25,000 cells (1µl of Transposomes) or 50,000 cells (2µl of Trnaposomes) as suggested [Buenrostro et al., 2013]. A sample with 1µl of Nextera TDE1 Enzyme obtained 11 cycles for amplification, whereas In-house Tn5 samples with 1µl and 2µl of Transposomes obtained 15 and 14 cycles respectively. Venn diagram indicates the overlapping regions for three libraries (Figure 1). Relevant sequencing statistics parameters are presented in Table 1.



Supplementary Figure 1 Venn diagram shows overlaps between regions covered with high-confident non-duplicated reads. Given numbers show the amount of sequence in Mb in the corresponding overlap.

Table 1 Alignment statistics.

Feature		Nextera (1µl)	In-house (1µl)	In-house (2µl)
Total number of reads		24676421	22027819	19304996
Number of reads aligned to reference (% of total)		24471978 (99.17)	21798080 (98.96)	19134929 (99.12)
Number of duplicates (% of total)		7042334 (28.54)	16291128 (73.96)	11013788 (57.05)
Number of high-confident* non-duplicated reads (% of total)		15411198 (62.45)	4485884 (20.36)	7048313 (36.51)
Mb of sequence covered with high-confident non-duplicated reads		615.12	189.83	274.13
Mb of sequence covered with N high-confident non-duplicated reads (% of total sequence covered with high-confident non-duplicated reads)	N=1	492.31 (80.04)	166.47 (87.69)	234.76 (85.64)
	2	82.41 (13.40)	12.45 (6.56)	21.97 (8.01)
	3	16.45 (2.67)	3.13 (1.65)	4.85 (1.77)
	4	5.89 (0.96)	1.72 (0.91)	2.47 (0.90)
	5	3.37 (0.55)	1.16 (0.61)	1.66 (0.61)
	6	2.35 (0.38)	0.85 (0.45)	1.23 (0.45)
	7	1.76 (0.29)	0.65 (0.34)	0.96 (0.35)
	8	1.39 (0.23)	0.51 (0.27)	0.77 (0.28)
	9	1.13 (0.18)	0.41 (0.22)	0.64 (0.23)
	10	0.95 (0.15)	0.34 (0.18)	0.54 (0.20)
	>10	7.11 (1.15)	2.14 (1.12)	4.28 (1.56)

*high-confident reads – reads with probability of wrong mapping lower than 0.05 according to their MAPQ score (MAPQ>13).