Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses

Melissa J. Fullwood, Chia-Lin Wei, Edison T. Liu, et al.

Genome Res. 2009 19: 521-532
Access the most recent version at doi:10.1101/gr.074906.107

Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses

Melissa J. Fullwood,[1,2] Chia-Lin Wei,[1] Edison T. Liu,[1,3] and Yijun Ruan[1,4,5]

[1]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore; [2]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore 117456, Singapore; [3]Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Singapore; [4]Department of Biological Sciences, National University of Singapore, Singapore 117456, Singapore

Comprehensive understanding of functional elements in the human genome will require thorough interrogation and comparison of individual human genomes and genomic structures. Such an endeavor will require improvements in the throughputs and costs of DNA sequencing. Next-generation sequencing platforms have impressively low costs and high throughputs but are limited by short read lengths. An immediate and widely recognized solution to this critical limitation is the paired-end tag (PET) sequencing for various applications, collectively called the PET sequencing strategy, in which short and paired tags are extracted from the ends of long DNA fragments for ultra-high-throughput sequencing. The PET sequences can be accurately mapped to the reference genome, thus demarcating the genomic boundaries of PET-represented DNA fragments and revealing the identities of the target DNA elements. PET protocols have been developed for the analyses of transcriptomes, transcription factor binding sites, epigenetic sites such as histone modification sites, and genome structures. The exclusive advantage of the PET technology is its ability to uncover linkages between the two ends of DNA fragments. Using this unique feature, unconventional fusion transcripts, genome structural variations, and even molecular interactions between distant genomic elements can be unraveled by PET analysis. Extensive use of PET data could lead to efficient assembly of individual human genomes, transcriptomes, and interactomes, enabling new biological and clinical insights. With its versatile and powerful nature for DNA analysis, the PET sequencing strategy has a bright future ahead.

Genomics holds much promise for huge improvements in human healthcare. However, genomics faces several practical challenges. Human genomes are read out as linear sequences, but in the cell, there are many complex interactions and mechanisms that operate around human DNA to transduce DNA information into biological function (The ENCODE Project Consortium 2007). Conventional DNA sequencing has been used to extensively explore genetic elements and structures; however, high sequencing costs and low throughputs have historically limited in-depth analysis of a broad range of genomic elements, making the development of new sequencing strategies necessary.

Next-generation sequencing technologies are transforming the field of genomic science (Schuster 2008). The currently available next-generation sequencing methods (Margulies et al. 2005; Shendure et al. 2005; Barski et al. 2007; Johnson et al. 2007) read DNA templates in a highly parallel manner to generate massive amounts of sequencing data, but the read length for each DNA template is short compared with that of traditionally used Sanger capillary sequencing instruments. This massively parallel and short read strategy of DNA sequencing opens many new ways for interrogating human genomes (Wold and Myers 2008). However, the short read lengths lead to serious limitations in applying this enormous sequencing power to many biological applications. Therefore, immediate efforts have concentrated on overcoming the limitation of short tags for genome-wide analysis.

The paired-end tag (PET) sequencing is one such strategy for improving DNA sequencing efficiency and enabling biological applications. In PET analysis, as outlined in Figure 1, short paired tags from the two ends of DNA fragments are extracted and covalently linked as ditag constructs for high-throughput sequencing and mapping to reference genomes, which demarcate the boundaries of the DNA elements in a genome landscape. PET analyses can use a variety of sources of nucleic acid: RNA, DNA, and subsets thereof enriched by molecularly manipulation protocols such as chromatin immunoprecipitation (ChIP). Hence, PET technology has many benefits (Box 1) that make it a unique fit to enhance the performance of next-generation sequencing technologies for genome function and variation analysis. Various applications of PET technology have shown immediate value by providing genome-wide and unique solutions for understanding genomes, transcriptomes, epigenomes, and interactomes. In the future, PET technology will continue to improve and expand to cover a greater range of applications in medical genomics. Eventually, it may help to overcome the challenges of personal genomics to make personal medicine a reality. Here, we provide a retrospective of the development of the PET sequencing strategy and its recent applications. We also discuss the challenges faced by PET technology and provide some perspectives.

The development of the PET strategy

The PET concept

The principal concept of the PET strategy is the extraction of only short tag signature information (20–30 base pairs) from the two ends of target DNA fragments, the pairing of the two tags for sequencing analysis, and then the mapping of the paired tag sequences to reference genomes for demarcating the boundaries of the target DNA fragments in the genome landscape.

[5]Corresponding author.
E-mail ruanyj@gis.a-star.edu.sg; fax 65-6478-9059.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.074906.107.
Freely available online through the *Genome Research* Open Access option.

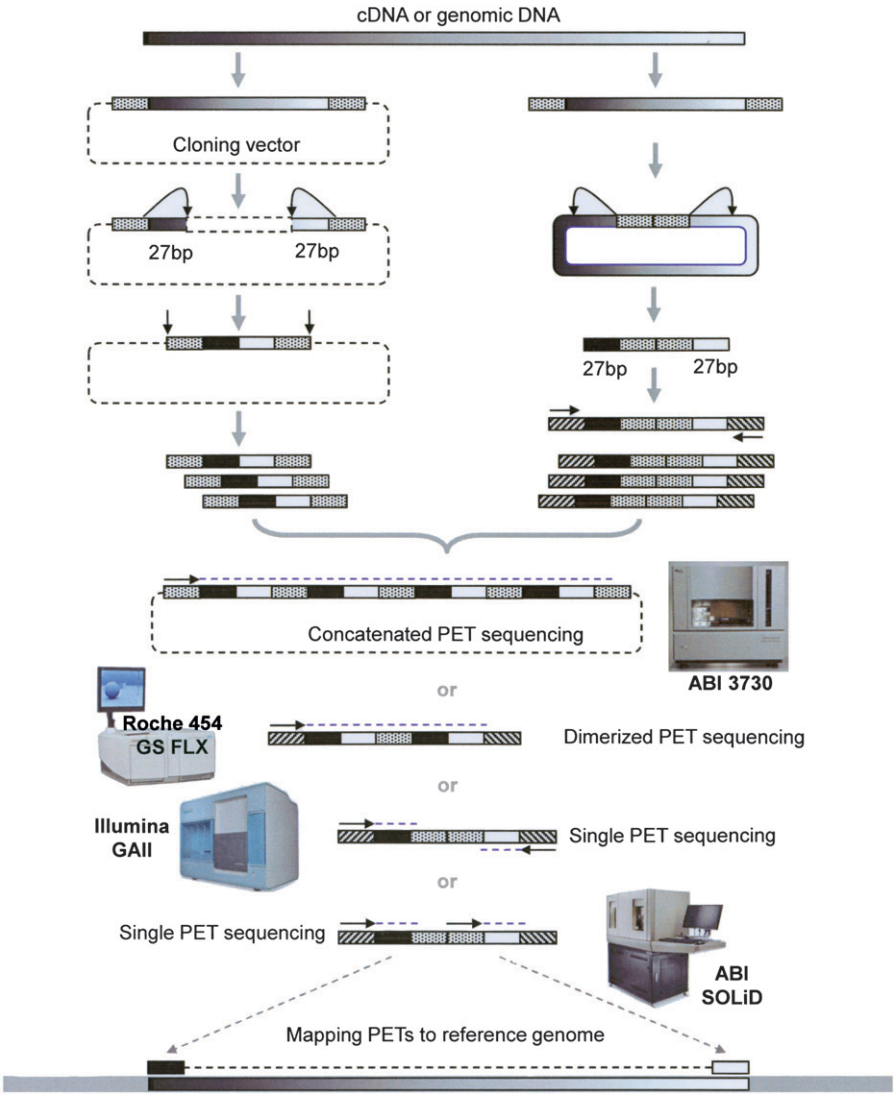

Figure 1. Schematic view of PET methodology. PET construction may be carried out through cloning-based or cloning-free procedures. In the cloning-based procedure, DNA fragments are ligated to cloning vector to introduce restriction sites such as EcoP15I to the 5′ and 3′ ends of insert DNA and link the two ends by a vector backbone. This is transformed into *E. coli* cells as a DNA library. EcoP15I digestion of the library will result in tag–vector–tag structures, which are re-ligated to form a single-PET library and further digested to release the PET constructs. In the cloning-free protocol, linker oligonucleotides containing EcoP15I sites are directly ligated to DNA fragments, followed by circularization, and then digestion by EcoP15I to release the PET constructs. The resulting PET constructs can be analyzed by concatemer sequencing using Sanger capillary instrument, dimerization sequencing using Roche 454 GS, paired-end sequencing using Illumina GA or Applied Biosystems SOLiD. The PET sequences are then mapped to reference genome sequences to demarcate the boundaries of the target DNA fragments.

The intellectual traces of the development of PET strategy converged from two important technological concepts: conventional paired end sequencing and short tag sequencing (Fig. 2). The first straightforward description of paired-end sequencing was reported by Hong (1981) using DNA inserts cloned into bacteriophage vectors and sequenced from both ends, thus reading twice as much sequencing data from long inserts. Then, in 1994, so-called "mate-pairs" of sequencing reads were used to help assemble the genome of *Haemophilus influenzae*, which was the first genome sequence of a free-living organism (Fleischmann et al. 1995). Turning to larger genomes, paired-end sequencing was an important component of early proposals (Venter et al. 1996; Weber and Myers 1997) and actual sequencing efforts for the *Drosophila* and the human genomes (Venter et al. 1998; Adams et al. 2000; Myers et al. 2000; Rubin and Lewis 2000; Lander et al. 2001). Later efforts to close gaps in assemblies also employed paired-end sequencing (Bovee et al. 2008). Besides the cost savings from two sequencing reads per template preparation rather than single reads, the distances between the two ends of size templates may be used to relate discrete contigs in assembling genomes. In addition, genomic regions containing repeats can be oriented and positioned by their connectivity to sequence specific regions offered by paired-end sequences.

The "chromosome jumping" method (Collins and Weissman 1984) was a novel approach that did not simply perform paired-end sequencing of an insert, but instead first cloned and enriched the junctions formed by circularized ligation of the two DNA ends of a large fragment, and then sequenced the junctions to reveal the two paired end sequences. "Chromosome jumping" was designed to enable big "jumps" of hundreds of kilobases span, as opposed to little "steps" by "chromosome walking," to aid in positional cloning of disease genes (Collins et al. 1987).

Around the same time, the short tag concept was developed to overcome the prohibitively high costs of sequencing. The underlying concept was that the nucleotide composition of a short DNA stretch is sufficiently specific to represent a longer DNA fragment. Expressed sequence tag (EST) was the first example of tag-based sequencing concept, by using single Sanger sequencing reads to tag cDNA sequences reverse transcribed from mRNA, instead of sequencing the full-length cDNAs (Milner and Sutcliffe 1983; Putney et al. 1983; Adams et al. 1991). Despite successful discovery of many genes (Adams et al. 1992), the high cost both in time and in resources for DNA sequencing promoted the desire to further shorten the sequenced tags, leading to the development of serial analysis of gene expression (SAGE) (Velculescu et al. 1995). In SAGE, a "tagging enzyme" (type IIS restriction endonuclease) was used to cut cDNA at a certain distance away from the restriction site introduced by adaptor sequence, and the short tags were concatenated for efficient sequencing analysis. Velculescu and colleagues demonstrated a conceptual breakthrough by showing that tags as short as 13 bp could be sufficient to match human cDNA sequences in existing databases. They then applied the SAGE approach to identify genes specific to cancer cells (Velculescu et al. 1995) as

Box 1. PET technology applications for the study of genomes and transcriptomes

Application	Benefits of PET	Techniques and references
Improve sequencing efficiency	PET template is compatible with next-generation machines Higher mapping specificity of PETs over single tags Decreased sequencing costs per template Retains information regarding the distance and relationship between the ends of DNA fragments	Paired-end ditag (PET) (Ng et al. 2005; Wei et al. 2006) Paired-end sequencing (PES) (Lander et al. 2001; Holt and Jones 2008) Paired-end mapping (PEM) (Korbel et al. 2007) Mate-pairs (Shendure et al. 2005) Ditag genome scanning (DGS) (Chen et al. 2008a) Paired-end genomic signature tags (PE-GST) (Dunn et al. 2007)
Transcriptome analyses	Identify 5′ and 3′ ends of transcription units Identify alternative TSSs and PASs Enables ultra-high-throughput genome-wide identification of gene fusion events, which is not possible with other methods	GIS-PET (Ng et al. 2005) GSC-PET (Carninci et al. 2005) RNA-PET and shotgun RNA-PET
TFBS and epigenetic site analyses Chromatin interaction analyses	Improved specificity and demarcation of fragments containing sites of interest Enable ultra-high-throughput, genome-wide, and de novo identification, which is not possible with other methods	ChIP-PET (Wei et al. 2006) PE-GST (Dunn et al. 2007) ChIA-PET
Genome structure variation analyses Genome assembly	Paired readout of DNA sequence for accurate genome assembly Span repeats and gaps Enable ultra-high-throughput genome-wide identification of small and large insertions, deletions and translocations, which is not possible with other methods	Ditag genome scanning (Chen et al. 2008a) PEM (Korbel et al. 2007) PES (Lander et al. 2001; Holt and Jones 2008) Mate-pairs (Shendure et al. 2005) DNA-PET

well as to characterize the yeast transcriptome (Velculescu et al. 1997).

Later, a new type IIS restriction enzyme, MmeI, was introduced. MmeI cuts DNA 18/20 bp downstream of its recognition site (Morgan et al. 2008). Use of MmeI enabled the development of LongSAGE to produce 20-bp tags (Saha et al. 2002). With such lengths, more LongSAGE tags could be specifically mapped to target transcripts and also be directly mapped to the reference genome for de novo identification of expressed genes, thereby making LongSAGE a valuable tool for annotating the then newly sequenced human genome using transcriptome data. Furthermore, the SuperSAGE method introduced EcoP15I, a type III restriction endonuclease that cuts 25/27 bp downstream of its recognition site, allowing for the extraction of even longer SAGE tags for higher mapping specificities (Matsumura et al. 2003). However, EcoP15I is problematic for SAGE protocols, because it requires two separated and inversely oriented recognition sites in supercoiled DNA and does not turn over (Raghavendra and Rao 2005). This special requirement is particularly suited to the double cleavage of the PET constructs. Furthermore, recent improvements have shown that the incorporation of sinefungin in EcoP15I reaction buffer allows cleavage at all recognition sites in a manner less dependent on DNA topology (Raghavendra and Rao 2005), which promises to make EcoP15I a useful laboratory tool.

Yet another short tag method that is based on the same set of concepts as SAGE analysis, but that uses a completely different sequencing strategy, is massively parallel signature sequencing (MPSS) (Brenner et al. 2000). Despite the novel and unconventional sequencing approach, the outcome of MPSS is the same as for SAGE.

SAGE and MPSS extract tags near the 3′ side of DNA fragments, often several hundred base pairs upstream from the 3′ ends of the cDNA. When mapped to the genome, such "internal" tags are often ambiguous in defining transcription units. Using just one point is insufficient to characterize a linear structure, but two end-points can accurately define a linear arrangement. The human genome sequence is a linear framework for identifying the complete contents of gene transcriptional units and provides an ideal target for tag-based approaches to demarcate the gene elements. An important conceptual advance after LongSAGE was the capture of immediate 5′ and 3′ short tags from cDNA fragments. To characterize 5′ transcription start sites (TSS) and hence identify gene promoters, cap analysis of gene expression (CAGE) was

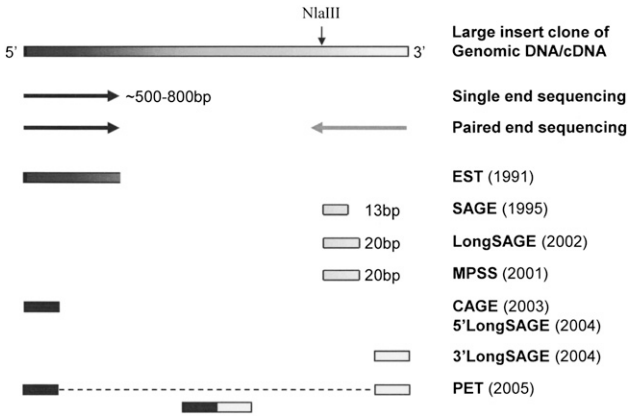

Figure 2. Sequencing-based methods for understanding genetic elements in genomes. The nucleotide information of DNA fragments can be spelled out by complete sequencing analysis. Alternatively, the large size of DNA fragments can be analyzed by single-end sequencing or paired-end sequencing. EST was the first tag-based approach, generating one tag per sequencing read, to represent full-length transcripts of expressed genes. The original SAGE tag extracts 13 bp next to the 3′ most NlaIII site to tag the transcripts. LongSAGE and MPSS use MmeI as the tagging enzyme to generate 20-bp tags that can be specifically aligned to reference genome sequences. The CAGE and 5′ LongSAGE tags are derived from the 5′ end of cDNA fragments, and the 3′ LongSAGE tags are derived from the 3′ end of cDNA fragments. PET covalently combines the 5′ and 3′ signature tags of the same DNA fragment into one ditag unit.

Box 2. Glossary related to PET technology

Glossary

Short Tag Sequencing: Sequencing only a short stretch of DNA information, typically less than 100 bp. The next generation sequencing platforms generate only short tag sequences, typically 16–50 bp. Ideally in the future, tags would be 50–100 bp.
EST (expressed sequence tag): Single Sanger sequencing reads from cDNA clone templates to tag expressed genes (Adams et al. 1992). ESTs are typically several hundred base pairs in length.
SAGE (serial analysis of gene expression): A method for preparation of short tags (13 bp), mostly for cDNA analysis to profile transcriptomes (Velculescu et al. 1995). Variants of SAGE include LongSAGE (20 bp) and SuperSAGE (25 bp).
MPSS (massive parallel signature sequencing): A short-tag approach similar to SAGE but using a ligation-based sequencing method to profile transcriptomes (Brenner et al. 2000).
CAGE (Cap-associated analysis of gene expression): A method using the Cap-trapper method to retain 5′ intact transcripts and a SAGE-like approach to extract 5′ tags (20 bp) for cDNA analysis and identify TSS (Shiraki et al. 2003). Variants include 5′ LongSAGE (Hashimoto et al. 2004; Wei et al. 2004).
PET (paired-end tag): Paired short tags are extracted from the ends of linear DNA fragments for ultra-high-throughput sequencing. The original method was Paired-End diTag (Ng et al. 2006). Variants of names include paired-end sequencing (PES), paired-end mapping (PEM), Mate-pairs, ditag genome scanning (DGS), and paired-end genomic signature tags (PE-GST) (Lander et al. 2001; Ng et al. 2005; Shendure et al. 2005; Dunn et al. 2007; Korbel et al. 2007; Campbell et al. 2008; Chen et al. 2008a; Holt and Jones 2008).
RNA-PET: PET applications to RNA or cDNA analysis for transcriptome profiling. An older variant is GIS-PET (Ng et al. 2005).
ChIP-PET (chromatin immunoprecipitation using PET): PET application to ChIP-enriched DNA fragments for identification of TFBS and epigenetic sites (Wei et al. 2006; Dunn et al. 2007).
ChIA-PET (chromatin interaction analysis using PET): PET application to cross-linked chromatin for identification of long-range chromatin interactions.
DNA-PET: PET application to genomic DNA analysis for the study of genome structural variations, and genome sequence assembly (Shendure et al. 2005; Korbel et al. 2007; Campbell et al. 2008).

introduced based on the Cap-trapper method (Carninci and Hayashizaki 1999) to retain 5′ intact transcripts, followed by "tagging" restriction digestion and the standard LongSAGE method to generate CAGE tags (Shiraki et al. 2003). Two other groups including us (Hashimoto et al. 2004; Wei et al. 2004) also independently developed similar approaches as 5′LongSAGE to map TSS. In addition, we simultaneously developed the companion 3′LongSAGE method, so as to map both 5′ TSS and the exact 3′ polyadenylation sites (PASs) to define the boundaries of expressed genes using two end tags (Wei et al. 2004).

Expanding from such a capacity, we then developed the paired-end ditag (PET) method that covalently links the 5′ tag and 3′ tag of a DNA fragment into a ditag structure for sequencing analysis (Ng et al. 2005), thus combining the benefits of the cost-effective SAGE and the linkage information from paired-end sequencing. We and others have since applied PET strategies to a variety of biological questions (Lander et al. 2001; Ng et al. 2005; Shendure et al. 2005; Wei et al. 2006; Dunn et al. 2007; Korbel et al. 2007; Campbell et al. 2008; Chen et al. 2008a).

The construction of PET structures

There are multiple methods for constructing PET structures (Fig. 1). The original PET method was "cloning-based," using plasmid vectors to link 5′ and 3′ tags. It was implemented as gene identification signature analysis using PETs (GIS-PET) for studying transcriptomes, in which the starting mRNA is converted into full-length cDNA (flcDNA) with flanking adaptor sequences containing MmeI restriction sites immediately next to both cDNA ends. The flcDNA fragments are then ligated to plasmids and transformed into *Escherichia coli* cells as a flcDNA library. The purified plasmids of the library are then digested with MmeI, which cuts into the cDNA insert to result in two 18/20-bp tags attached to the vector backbone. The tag–vector–tag structures are recircularized under intramolecular ligation conditions so that the two tags are joined covalently. The resulting single PET library can be amplified in bacteria cells, and the PET constructs are then excised by a restriction digestion from purified PET library plasmids (Ng et al. 2007).

A recent alternative for PET construction involves direct circularization of the target DNA fragments with linker oligonucleotides that covalently join the two ends of a DNA fragment. As the linker sequence is typically designed to contain two MmeI or EcoP15I sites flanking the two ends of the circularized DNA fragment, restriction digestion with these enzymes would release the tag–linker–tag structure for sequencing. This strategy was first demonstrated in resequencing an *E. coli* genome using the polony sequencing method (Shendure et al. 2005). Besides tagging enzymes such as MmeI and EcoP15I that generate uniform sizes (18/20 bp and 25/27 bp) of PET constructs for easy manipulation, frequently cutting restriction enzymes and physical shearing by nebulization are also choices for generating randomly sized tag–linker–tag constructs. As reported (Korbel et al. 2007), circularized DNA was randomly sheared by nebulization, and the fragments with biotinylated linkers were isolated using streptavidin. This method produces tags with a median size of 106 bp and is very useful for obtaining long tags because no type IIS or III restriction enzyme is currently known to produce tags more than 30 bp; however, many PETs prepared this way are unbalanced with tags of lengths under 15 bp, which would mean that these sequences would have to be discarded.

A benefit of the cloning-based method is that it preserves the original full-length cDNA or ChIP DNA fragments in a sustainable format of library clones. However, the construction process is long (2–4 wk) and can be technically challenging. By contrast, the cloning-free method is rather straightforward and can avoid many biases related to cloning.

Sequencing analysis of PET constructs

PET constructs can be sequenced by all available sequencing platforms (Fig. 1). Before the arrival of next-generation DNA sequencing instruments, the traditional method for short tag sequencing was to concatenate the tags into long stretches of DNA for Sanger sequencing. An average sequencing read would yield 20–30 tags. This concatenation sequencing strategy was applied to PET sequencing with great success, demonstrating the value of PETs for transcriptome analysis (Ng et al. 2005) and genome functional analysis (Loh et al. 2006; Wei et al. 2006).

The short templates of PET constructs are ideally suited to analysis by next-generation DNA sequencing methods that are massively parallel but have short read lengths (next-generation DNA sequencing has been reviewed in detail; Holt and Jones 2008). One of the first successful next-generation sequencing methods, the Roche 454 GS20, was published in 2005 (Margulies et al. 2005). We

conceived of a dimerization method to ligate two units of PET together to form a diPET template that is ~80 bp, perfectly fitting within the read length of the GS20 pyrosequencer (then 100 bp). Using this approach, a single run of diPET templates can generate a half million PET sequences (Ng et al. 2006). This advance represented an immediate 100-fold increase in efficiency for PET sequencing when compared with the use of the Sanger sequencing method. In addition, further PET ligations can create longer length-controlled templates, allowing for scalability as sequencing read lengths increase.

Toward the end of 2006, the Illumina Genome Analyzer (GA) sequencing machine was introduced to the market. The most impressive feature is its massively parallel capacity for reading up to 80 million DNA templates simultaneously, even though it reads only ~36–50 bp from each template (Barski et al. 2007; Johnson et al. 2007; Mikkelsen et al. 2007). The present GAII system has three ways to generate PET sequences. First, PET constructs made from long DNA fragments (cDNA or genomic DNA) can be read from both directions. In this approach, single-strand DNA templates are prepared on the flow cell surface for the first-strand DNA read from one direction. After that, the second-strand DNA is synthesized in situ to replace the first one and then read from another direction. The second approach is simply to sequence the entire PET construct, since the GAII model can read more than 75 bp. The third approach is to bypass PET construction and simply sequence the two ends of DNA fragments using the paired-end sequencing method described above. Although this last approach is simple and straightforward, it is limited to short DNA fragments that can be amplified by bridging PCR on the surface of the flow cells, and therefore, only short fragments of a few hundred base pairs can be paired-end sequenced by this method. For longer DNA fragments, paired-end sequencing has to be done through making PET libraries first.

SOLiD is another massively parallel short-tag sequencing platform introduced in late 2007 by Applied Biosystems. This platform was adapted from the polony sequencing method (Shendure et al. 2005). Bearing the limits of short tags in mind, the current version of SOLiD is designed mainly for paired-end sequencing and can read about 200 million tags for 25 bp from each end per machine run of two weeks. In this system, the first tag is read using a primer that primes onto the flanking adaptor, while the second tag is sequenced from the middle linker of the PET construct.

All three currently available next-generation technologies have advantages and disadvantages for PET sequencing analysis: The 454 Life Sciences (Roche) system is the most versatile and has the fastest turnover time, but the reagent cost for running it is relatively expensive. The Illumina GAII is very robust, cheap, and high-throughput, but paired-end sequencing takes up to a week of running time. Applied Biosystem's SOLiD is cheap and probably has the highest throughput but also the slowest runs. Additional advancements in current and future next-generation sequencing machines promise to bring further improvements in costs, read lengths, throughputs, preparation times, run times, and accuracies (Metzker 2005). Examples include Helicos (Harris et al. 2008) and Pacific Biosciences (Eid et al. 2008) for single molecule sequencing, which could result in lower costs, higher throughput, and less sample input required for sequencing. The Helicos system may significantly increase the output of tag-based sequencing capacity up to billions of tags per sequencing slide, and its sequencing methods can be adapted to PET analysis. The Pacific Biosciences method has been developed toward the readout of long DNA circular templates. Its current capability may read a few thousand base pairs. Therefore, PET protocols that can create circular DNA molecules may be readily adapted to this method.

After sequencing, PETs are mapped to the reference genome. The large volume of PET sequences generated from each machine run have imposed immense challenges on how to efficiently process the data and accurately map the PET sequences to reference genomes. An example of a PET processing solution is provided by PET-Tool, a user-friendly software package that does all steps, including PET extraction from raw sequence reads, PET mapping to reference genomes, and data management for hosting different PET experiments (Chiu et al. 2006). PET-Tool has since been updated to accommodate next-generation sequencing platforms. Other examples have been described and reviewed in the literature (Li et al. 2008a,b; Pop and Salzberg 2008; Zerbino and Birney 2008).

Applications of PET technology

PET technology is superior to single-tag sequencing for genome structure and function analysis and remains versatile, such that different forms of nucleic acids can be analyzed in different PET applications. PET can be applied to RNA (RNA-PET) for transcriptome analysis; to DNA (DNA-PET) for genome structure variation and aid genome sequence assembly; to manipulated DNA fragments such as ChIP-enriched DNA (ChIP-PET) for mapping transcription factor binding sites (TFBSs); and to proximity-ligated DNA for chromatin interaction analyses (ChIA-PET) (Fig. 3). These applications have been used and will be used to generate many more whole-genome maps as rich data resources to annotate the genomes. In the following sections, we review the applications of the PET technology in genome analysis and future perspectives.

RNA-PET for transcriptome studies

Transcriptome studies include understanding gene structures and transcription dynamics (Fig. 3). The structural elements of genes include exons, introns, TSSs, and PASs. The gold standard for uncovering gene structure is flcDNA sequencing (Carninci and Hayashizaki 1999). However, this is a very expensive and laborious approach. Whole-genome tiling arrays have proved effective for identifying exons and measuring transcription dynamics (Kapranov et al. 2002; The ENCODE Project Consortium 2007); however, arrays can be ambiguous in defining the exact boundaries of transcription units particularly in gene dense regions, because array data lack connectivity information between exons. Single-tag based approaches are only effective in defining TSSs and quantifying alternative usage (Shiraki et al. 2003; Hashimoto et al. 2004). Recently, shotgun sequencing of transcripts (RNA-seq) by Illumina GA and Applied Biosystems SOLiD has been used to profile genes and has generated an unprecedented wealth of gene information, particularly with regard to new exons and possible alternative splicing forms (Marioni et al. 2008; Morin et al. 2008; Mortazavi et al. 2008; Nagalakshmi et al. 2008; Sultan et al. 2008; Wilhelm et al. 2008). However, like tiling arrays, RNA-seq data do not provide connectivity between exons of transcription units.

By contrast, using the PET approach to sequence RNA, RNA-PET data would demarcate the first and the last exons, so as to define the TSSs and PASs, as well as the connectivity between the two sites. However, an obvious limitation is that RNA-PET will not reveal information regarding internal exons. Therefore, RNA-PET is a complementary approach to tiling array and RNA-seq data. A unique feature that sets RNA-PET apart from other methods is its ability to detect unconventional fusion genes.

The early version of RNA-PET was a cloning-based method, GIS-PET analysis (Ng et al. 2005). In GIS-PET, flcDNA is prepared using the

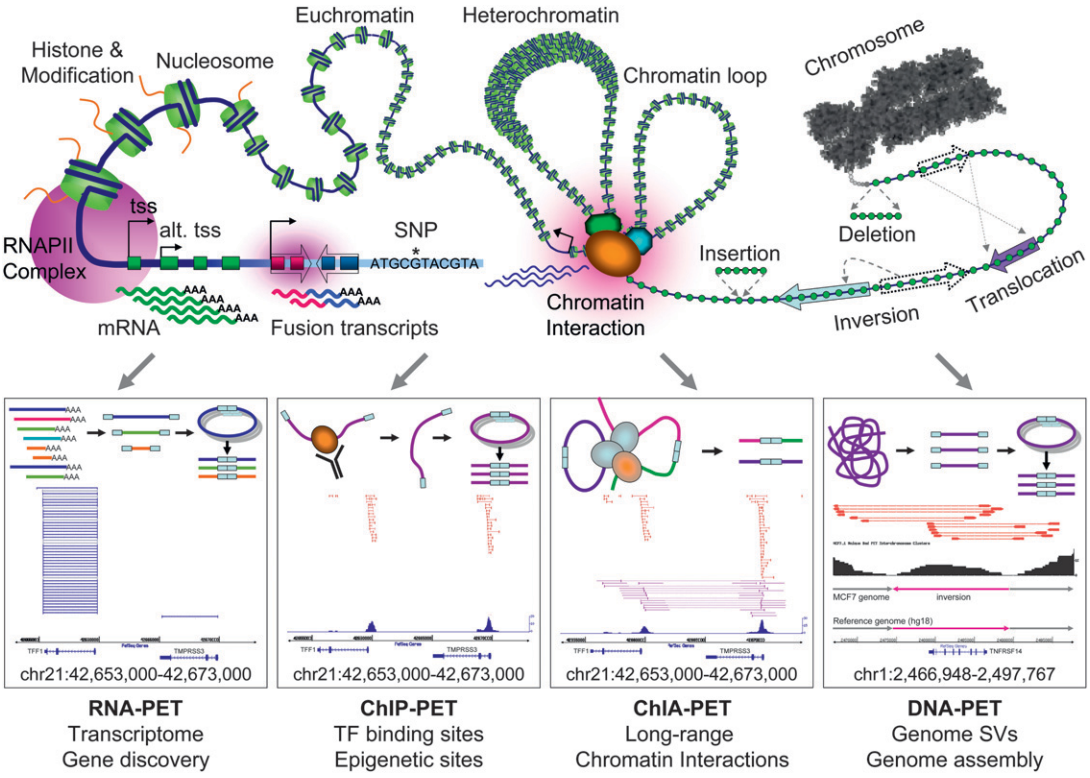

Figure 3. PET applications to address genome biology questions. Cells have many different mechanisms for processing, modifying, controlling, and transducing information encoded in the genome. The PET technology can be applied to investigate many questions regarding nuclear processes, such as transcriptomes by RNA-PET, transcription and epigenetic regulation by ChIP-PET and ChIA-PET, as well as genome structural variation by DNA-PET. Examples of PET data from GIS-PET (an early version of RNA-PET), ChIP-PET, and ChIA-PET experiments of human breast cancer MCF-7 cells with estrogen induction treatment at the *TFF1* locus (chr21:42,653,000-42,673,000) are shown: the high level of *TFF1* gene expression and the low level of *TMPRSS3* gene expression; the ERα binding at the *TFF1* promoter and enhancer sites; and the long-range chromatin interactions between the two ERα binding sites. An example of DNA-PET data at the *TNFRSF14* locus in the genome of MCF-7 cells shows an inversion event detected by two clusters of discordant DNA-PET cluster mapping.

PET method: The capped 5′ ends and the poly(A)-tailed 3′ ends are captured in a pairwise manner by 20-bp signature tags, and these paired-end sequences may then be mapped to the genome. The flcDNA library can also be normalized before PET sequencing analysis, thus enriching for rarer clones, and hence allowing for more efficient discovery of low-abundance genes (Carninci et al. 2005).

GIS-PET has been applied to the studies of transcriptome in mouse embryonic stem cells (Ng et al. 2005), various mouse tissues as part of the FANTOM3 project (Carninci et al. 2005), and a number of human cells as part of the ENCODE project (The ENCODE Project Consortium 2007). Many isoforms of transcripts with alternative TSSs and PASs were characterized, and large numbers of novel transcription units were identified. In mouse ES cells, we found and validated a *trans*-splicing fusion mRNA between *Ppp2r4* and *Set*, in which the first exon of *Ppp2r4* was joined to the second exon of *Set*. Upon further characterization, we found that this fusion gene is preferentially expressed in embryonic as opposed to adult tissues, and the fusion gene might encode a new functional protein, suggesting that the fusion might play a role in early development in mice (Ng et al. 2005).

Human cancer cell lines are known to contain extensive chromosomal aberrations. Fusion genes created through chromosomal rearrangements could play roles in oncogenesis. Several successful diagnostic methods and therapies target fusion gene products (Mitelman et al. 2007); for example, Gleevec targets the *BCR/ABL* fusion in chronic myelogenous leukemia (Mauro et al. 2002). We applied GIS-PET to two human cancer cell lines to understand unconventional fusion transcripts (Ruan et al. 2007). From an analysis of 865,000 GIS-PETs from MCF-7 and HCT-116, we found 70 fusion genes, including a fusion between *BCAS3* and *BCAS4* that had been previously identified in MCF-7 cells. Other fusion genes identified and validated by RT-PCR included *CXorf15/SYAP1* and *RPS6KB1/TMEM49*. Interestingly, *SYAP1* has been implicated in chemotherapy response (Al-Dhaheri et al. 2006), and *RPS6KB1* is an oncogenic marker (van der Hage et al. 2004), suggesting a possible role for these fusion genes in cancer progression.

These GIS-PET studies showed that a unique capability of the PET sequencing strategy is high-throughput identification of fusion genes. We are now developing a cloning-free full-length RNA-PET method that bypasses the cloning steps, directly introduces the adaptor sequences to the two ends of flcDNA fragments, and then circularizes the DNA for PET analysis using next-generation sequencing platforms (Fig. 1). An even more straightforward RNA-PET strategy would be to simply perform shotgun paired-end sequencing of cDNA templates that have been fragmented to a few hundred base pairs. In theory, this shotgun RNA-PET approach should be able to identify the junction points of potential fusion transcripts if the sequence coverage is deep enough. The shotgun RNA-PET strategy can be applied to poly(A)+, poly(A)−, and noncoding RNA.

In conclusion, RNA-PET is the most efficient and accurate approach to demarcate the boundaries of transcription units of genes and complements other methods for transcriptome studies. The most unique benefit of RNA-PET is the ability to identify unconventional fusion transcripts. A large-scale RNA-PET program to investigate fusion genes could lead to discoveries of new candidate biomarkers for diagnostic and therapeutic options.

ChIP-PET for identifying regulatory and epigenetic elements

Besides gene coding sequences, genomes contain many noncoding elements that have important regulatory functions through interaction with protein factors (Fig. 3). Thus, mapping protein factor binding sites in the genome is an important starting point for understanding regulatory circuits. The traditional mainstream approach for mapping such protein/DNA interactions is ChIP-chip, a method in which ChIP-enriched DNA fragments are detected by whole-genome microarray (chip) hybridization (Ren et al. 2000).

ChIP-enriched DNA fragments can also be analyzed by sequencing approaches. ChIP analysis using PET sequencing (ChIP-PET) represents one of the first serious sequencing-based approaches to characterize ChIP-enriched DNA fragments (Wei et al. 2006) (Fig. 3). ChIP-PET provides linked 5′ and 3′ sequences for ChIP-enriched DNA molecules, which are mapped to the reference genome such that the complete ChIP DNA fragment can be inferred from the genome sequence in between the 5′ and 3′ tags, and the enriched TFBS can be determined.

We applied the ChIP-PET method to examine TP53 (also known as p53) transcription factor binding sites in HCT116 colon cancer cells and found 542 high confidence binding sites (Wei et al. 2006). Over 99% of these high confidence binding sites could be verified by ChIP-qPCR validation experiments, and PET-defined binding regions could be narrowed down to as little as 10 bp. We further demonstrated that these binding sites were likely to be functional by showing their clinical relevance to TP53-dependent pathways in primary cancer samples. Interestingly, we discovered that in addition to 5′ promoter proximal regions of genes, we could find many distal TFBSs that were far away from gene promoters. We went on to use ChIP-PET to map whole-genome binding profiles for a number of important transcription factors, including POU5F1 (also known as OCT4) and NANOG (Loh et al. 2006); MYC (Zeller et al. 2006); ESR1 (also known as ERα) (Lin et al. 2007); and NFKB (Lim et al. 2007). We also applied ChIP-PET to map epigenetic marks for epigenomic profiles of histone modifications in human embryonic stem cells (Zhao et al. 2007). Variants of ChIP-PET have also been reported, such as paired-end genomic signature tags (PEGST), which has been used to identify transcription factor binding sites and DNA methylation patterns (Dunn et al. 2007). ChIP-chip and ChIP-PET methods agree well on strong binding sites (Hudson and Snyder 2006; Euskirchen et al. 2007).

The arrival of next-generation sequencing is critical to further advance the sequencing-based measurement of ChIP-enriched DNA. We first incorporated the Roche 454 sequencing platform (GS20 and GS FLX) for ChIP-PET sequencing (Ng et al. 2006) and applied this approach to characterize the epigenomic profiles of histone modifications in mouse embryonic stem cells (Zhao et al. 2007). Recently, the ChIP sequencing strategy has been further extended by taking advantage of the Illumina GA sequencing platform. In the ChIP-seq method, randomly sheared ChIP DNA is ligated to adaptors and amplified by PCR. A narrow size range of the amplicon (200–300 bp) is analyzed by single direction Illumina GA sequencing. Many ChIP experiments yield very little DNA, therefore the low sample amount requirements (10 ng) for the Illumina GA instrument, combined with high-throughput and low cost, make this option very attractive. ChIP-seq has been used to generate exciting results in mapping histone modifications and TFBSs (Barski et al. 2007; Johnson et al. 2007; Chen et al. 2008b). Even more recently, Illumina has developed a paired-end sequencing method, which can be used to sequence the two ends of ChIP DNA fragments, instead of only single reads.

With the eminent success of ChIP-seq (single-end read of ChIP DNA) for identifying TFBSs, it is debatable whether ChIP-PET (paired-end read of ChIP DNA) is necessary. The benefit of ChIP-PET over ChIP-seq is that it provides two connected DNA end tags for unambiguous identification of TFBS locations. Many validated TFBSs have been found to be present in repeat regions (Euskirchen et al. 2007), for reasons related to the evolution of TFBSs through association with repeats (Bourque et al. 2008). Single-end reads in repeat regions can be ambiguous and are often discarded for further analysis. However, paired-end reads may span into repeat regions for accurate mapping. Hence, these additional tags are necessary for precise identification of an important class of TFBSs, as well as truly unbiased global TFBS sequencing.

Collectively, ChIP-PET and ChIP-seq powered by Illumina and other massively parallel tag sequencing platforms have generated and will continue to generate valuable maps of protein factors interacting with genomic DNA in the genomic landscape. From these analyses, general pictures of transcription factor binding have started to emerge. Many transcription factors show complex binding patterns with relation to target genes including TP53 (Wei et al. 2006), POU5F1 and NANOG (Loh et al. 2006), and others. Many TFBSs are far away from promoters of target genes. How remote regulatory elements function, if at all, is still largely unknown.

ChIA-PET for identifying chromatin interactions

The applications described above have concentrated on finding genetic elements in linear DNA. However, thinking of genomic information in a one-dimensional form is far less than sufficient to elucidate the complexity of genome functions implemented through three-dimensional organization in limited nuclear space. Evidence suggests that DNA molecules are packaged with protein factors to form chromatin fibers and folded into higher-order structures and eventually chromosomes as organizational units (Woodcock 2006). Genetic elements may interact by coming into close proximity as a result of chromosome conformation to produce spatial-based functions (Fig. 3). Genome functions such as transcription and replication could be closely associated with higher-order genome topology (Fraser and Bickmore 2007); however, we are still in the early stages of understanding the complex of structure–function interplay in the human genome.

Much of our current understanding of genome organization and function has come from two categories of technologies: molecular probing and molecular interaction mapping. Molecular probing is to visualize the three-dimensional structure of genome organization in nuclear compartments and monitor the dynamics in living cells. Electron microscopy and atomic force microscopy have been used to visualize DNA loops with limited success (Mastrangelo et al. 1991; Yoshimura et al. 2004). Fluorescence in situ hybridization (FISH) and Cryo-FISH using fluorescently labeled DNA or RNA probes to visualize specific regions of chromatin have been used to generate much valuable information regarding long-range interactions and chromatin conformation in

the entire nucleus (Cremer and Cremer 2001; Osborne et al. 2004; Branco and Pombo 2006). However, FISH-based approaches are limited by low resolution and are incapable of studying multiple loci at the same time.

Molecular interaction mapping approaches identify functional DNA elements that are in close spatial proximity and hence are likely to be potential interaction points in spatial genomic organization. One of the first experiments in this area was the nuclear ligation assay (Cullen et al. 1993), which sought to understand the potential of enhancer sites to form looping interactions. The enhancer sites were cloned into minichromosomes that were stably transfected into a rat cell line. The chromatin was then digested with restriction enzymes and ligated under dilute conditions to join the sticky ends. This ligation product was then inspected using PCR with specific primers for the presence of particular known interaction sites that bring together target genomic regions and the transfected minichromosomal regions. This nuclear ligation approach was further optimized in the Chromosome Conformation Capture (3C) protocol (Dekker et al. 2002), which was the first application to investigate in vivo chromatin interactions in yeast cells without exogenous DNA sequences. In 3C, chromatin is formaldehyde cross-linked and restriction enzyme-digested, the tethered DNA fragments are ligated in a dilute manner and reverse cross-linked, and the ligation products are detected by PCR similar to the nuclear ligation assay. 3C was subsequently applied to the study of long-range chromatin interactions between the beta-globin locus and locus control regions (LCRs) in mammalian cells (Tolhuis et al. 2002). Further, the 3C method had been combined with ChIP in the ChIP-loop assay to identify long-range interactions mediated by MECP2 at the *Dlx5-Dlx6* locus (Horike et al. 2005). However, 3C or ChIP-3C methods are limited to the detection of specific interactions using prior knowledge or perception of the existence of such interactions. To overcome this limitation, a number of groups have developed associated chromatin trap (ACT) (Ling et al. 2006), Chromosome Conformation Capture using Chip (4C) (Simonis et al. 2006), Circular Chromosome Conformation Capture (also called 4C) (Zhao et al. 2006), Open-ended Chromosome Conformation Capture (Wurtele and Chartrand 2006), and Chromosome Conformation Capture Carbon Copy (5C) (Dostie et al. 2006) methods to expand the scope of detection for chromatin interactions. However, all these methods are essentially extensions of 3C using target-specific PCR assays to detect interactions of known target with unknown regions. Although the current methodologies are valuable for providing insights of chromatin interactions at limited loci or limited resolutions, they are constrained by their inability to provide a whole-genome view of chromatin interactions.

We have previously applied the PET approach to identify unconventional fusion genes through mapping of the 5′ exon of one gene in a genomic locus and the 3′ exon of another gene in a different genomic location. The same concept can also be extended to characterize artificially fused DNA fragments, such as nuclear proximity ligation products. With this in mind, we propose a new strategy for whole-genome chromatin interaction analysis using paired-end tag sequencing (ChIA-PET) (Fig. 3). The basic concept of ChIA-PET is to introduce a linker sequence in the junction of two DNA fragments during nuclear proximity ligation to build connectivity of DNA fragments that are tethered together by protein factors. Therefore, all linker connected ligation products can be extracted for the tag–linker–tag constructs that can be analyzed by ultra-high-throughput PET sequencing. When mapped to the reference genome, the ChIA-PET sequences are read out to detect the relationship of two DNA fragments in chromatin interactions captured by chromatin proximity ligation. As this strategy is not dependent on any specific sites for detection like 3C or 4C, ChIA-PET has the potential to be an unbiased genome-wide approach for de novo detection of chromatin interactions.

We anticipate that the complexity of potential substance for proximity ligation is high, the nonspecific noise can be excessive; hence, the cost of sequencing such material to the required depth to find true proximity ligation products can be prohibitive even for the most advanced sequencing technology currently available. To reduce the complexity and background level, we propose to use ChIP against specific protein factors to enrich the corresponding chromatin fragments before proximity ligation. This enrichment approach would not only make the ChIA-PET sequencing practical by reducing the complexity but also add specificity to the identified interaction points. Depending on the protein factors used for ChIP enrichment, ChIA-PET analysis can be applied to the detection of all chromatin interactions involved in a particular nuclear process. For instance, the use of general transcription factors or RNA polymerase II components would identify all chromatin interactions involved in transcription regulation; the use of protein factors involved in DNA replication or chromatin structure would allow identification of all chromatin interactions due to DNA replication and chromatin structural modification. More specifically, the use of specific transcription factors for ChIA-PET analysis would further reduce library complexity and add specificity, and therefore, enable examination of specific chromatin interactions mediated by particular transcription factors. Our preliminary experimental data have demonstrated that ChIA-PET can generate PET sequences that identify TFBS and interactions between remote binding sites. With further development and optimization of the ChIA-PET prototype protocol, we expect that this whole-genome approach for unbiased and de novo discovery of long-range chromatin interactions will help to establish an emerging field for studying genome interaction and regulation networks in three dimensions.

DNA-PET for genome structure analysis

Genomes are variable at both the base-pair level and large-scale structural levels (Fig. 3). Genome variations at nucleotide level such as single-nucleotide polymorphisms (SNPs) and mutations are well understood to have functional roles in normal traits and diseases (Shastry 2007). However, our understanding of large structural rearrangements in the human genomes is just beginning. SAGE-based digital karyotyping (Dunn et al. 2002; Wang et al. 2002), array comparative genomic hybridization (aCGH) (Pinkel et al. 1998), and whole-genome tiling arrays (Kim et al. 2005) have contributed to this field by identifying large chunks of deletions and assessing copy number variations of amplified regions in disease genomes compared to normal and reference genomes. However, neither the single-tag sequencing approach nor the hybridization methods can identify balanced structural variations such as insertions, inversions, and translocations in genome rearrangements. Although paired-end sequencing of large genomic DNA inserts in fosmid and bacterial artificial chromosome (BAC) clones using conventional sequencing technique has generated highly valuable information regarding human genome structural variations (Tuzun et al. 2005; Kidd et al. 2008), the cost of such efforts is prohibitive.

PET sequencing of genomic DNA fragments (DNA-PET) is an ideal method for sequencing and assembling genomes as well as

studying genome structural variations (Korbel et al. 2007). DNA-PET provides linked 5′ and 3′ tag sequences from genomic DNA fragments of specific sizes, for example, 400 bp (Campbell et al. 2008) or 3 kb (Korbel et al. 2007). To accomplish this, genomic DNA is sheared by nebulization and purified in specific size range. PET constructs are then obtained from the genomic DNA fragments, followed by sequencing and mapping to the reference genome to infer the size of DNA fragments. Most DNA-PET sequences are concordant to the reference genome with correct orientation and specific size range. DNA-PETs with discordant mapping orientation and distance would be located at the breakpoints of structural variations between the reference and the test genomes.

The DNA-PET method was first demonstrated in resequencing an evolved *E. coli* genome using the polony sequencing-by-ligation method (Shendure et al. 2005). In the effort to study human genomic structural variation (Korbel et al. 2007), genomic DNA from an African and a European individual were sheared into 3-kb fragments and DNA-PETs of the fragments were sequenced by Roche 454 instrument and mapped to the human reference genome. Simple deletions were predicted from DNA-PET sequences mapped spans that were much larger than 3 kb, and simple insertions were predicted from those much shorter than 3 kb, while inversions were predicted from altered end orientations. More complex structural variations were also found from PET mapping patterns that were discordant. Through this analysis, 1297 structural variations were found. Forty-five percent of structural variations were shared between the two individuals, suggesting that some structural variations might be common. Hotspots of structural variations were found, which turned out to be regions that have been found to be involved in genomic disorders. Additionally, many structural variations could affect gene functioning by removing exons, creating gene fusions, being present in introns, altering gene orientation, or amplifying the genes. Interestingly, genes with protein products that were associated with interactions with the environment contained more structural variants than expected by chance (Korbel et al. 2007). This observation suggests a possible role for differences in these genes in order to cope with differences in environments. As an alternative to random shearing, restriction digestion to fragmentize genomic DNA for DNA-PET analysis was also tried (Chen et al. 2008a). In this study, human leukemia Kasumi-1 DNA and a normal control were tested, and structural variations between the test genomes and the reference human genome sequences were identified.

The DNA-PET approach has also been applied to map cancer genome rearrangements (Campbell et al. 2008). The authors took an even simpler approach to generate PET sequences from two cancer cell line genomes; genomic DNA was sheared to an average size of 200 bp and isolated, and 29–36 bp at either end were sequenced by the Illumina paired-end sequencing method. About 7 million DNA-PET sequences from each of the two cell lines were uniquely mapped to the reference genome and more than 400 rearrangements were identified to base-pair resolution. Because of the high density of the tag sequence data, accurate copy numbers of amplified regions in the human cancer genome were also obtained. Further analysis of the data allowed the authors to identify 103 somatic rearrangements and 306 germline structural variations. This suggests that many somatic variations are associated with amplicon regions of the genome, while most germline rearrangements are mediated by retrotransposition elements such as *Alu*Y and LINE. This work demonstrates the feasibility of systematic genome-wide efforts to characterize the architecture of complex human cancer genomes. It should be noted that the distance between the PETs in this situation was too short to span repeats; however, the benefits of this method are that it is highly economical and easy to prepare. It should also be noted that the authors had to discard 48% of the sequenced reads as they did not map to the reference genome. These results suggest that inefficiencies in the library construction steps or the new Illumina paired-end sequencing method reduced the amount of data that might otherwise have been obtained from the sequencing run. Moreover, of the reads that did map well, the authors excluded 38% because they precisely duplicated other sequences from the same library. The authors suggest that these sequences might have been preferentially amplified during the PCR step. Increased amounts of starting genomic DNA, reduction in the number of PCR cycles used, and PCR amplification of the entire ligation mix as opposed to a small aliquot are measures that could increase the complexity of the resulting library. In addition, care should be taken during library preparation such that all steps go to completion, to ensure that the resulting library is of high quality.

The power of connectivity provided by DNA-PET may be used to facilitate the assembly of whole-genome shotgun sequence reads for de novo genome sequencing and resequencing. With the current dramatic increase of DNA sequencing capacity, getting enough coverage of shotgun reads is no longer a serious issue. Using the massively parallel short tag sequencing platforms, 10× to 30× base-pair coverage of a human genome can be generated with a reachable budget and within months. However, assembling such short tag sequences alone would result in large numbers of contigs that cannot be joined up with each other. The real challenge is how to connect and orientate these contigs into the complete assembly of a complex genome such as the human genome. DNA-PET experiments (Korbel et al. 2007; Campbell et al. 2008) and computer simulations (Shendure et al. 2005) suggest that PET sequences could be useful for de novo complex genome sequencing.

A critical aspect in developing such a DNA-PET based strategy is the construction of PETs for large DNA insert fragments, such as 10-kb or even 100-kb fragments. One reason for this is that mammalian genomes have many repeat elements that are greater than 3 kb long. DNA-PETs that are longer than the length of repeat regions are needed to assemble chromosomes, by crossing over the repeated regions. Another reason is that longer DNA fragments will enable the discovery of insertions and intrachromosomal translocation events greater than 3 kb, which is the upper limit of the current DNA-PET approaches. In our laboratory, we are able to generate DNA-PET sequences up to 15 kb genomic DNA inserts. Our preliminary data show that large insert DNA-PET is clearly better than short insert DNA-PET, because large insert DNA-PET gives higher physical coverage. In silico analyses support this finding: As the length of the insert DNA increases, the physical coverage increases, and hence the probability of detecting a fusion point increases (Bashir et al. 2008). With these improvements, the DNA-PET method combined with ultra-high-throughput sequencing platforms will become a very powerful strategy for de novo genome sequencing and individual genome resequencing. Just as the human genome sequencing experiments were performed with paired-end sequences from inserts of multiple sizes, a combination of multiple DNA-PET sizes could be useful in individual human genome resequencing and de novo sequencing. Small structural variants might be detected and small repeats might be crossed using 1-kb to10-kb DNA-PET approaches, and large structural variants might be detected and large repeats might be crossed using 100-kb DNA-PET approaches. If this strategy proves successful,

this development in DNA-PET will pave the way for personal genomic approaches to resequence many individual human genomes. In conclusion, the DNA-PET strategy for genome structure analysis has immediate value and long-term promise. Already, DNA-PET with the current sequencing capacity can provide comprehensive characterizations of human structural variations associated with genetic diseases. Further development of DNA-PET with improved speeds, reduced costs, and the ability to use clinical samples would create a new karyogenomics platform for clinical implementation. In the long term, DNA-PET can become a vital part in the concept of personal genomics for personal medicine.

The future of PET technology

The unique feature of building connectivity between two points of DNA from linear and nonlinear structures in PET analysis has tremendous value in many aspects of genomic analysis that cannot be simply and easily replaced by just improving sequencing capacity in the near future. The PET concept is versatile, allowing for ready adaptation to new sequencing technologies. In the future, PET technology will grow by incorporating new sequencing technologies, overcoming existing limitations, and finding new applications for answering biological questions.

One limiting factor for PET analysis compared with single-tag sequencing such as RNA-seq and DNA-seq is that PET requires relatively more starting samples. This is because PET protocols involve more molecular manipulations and at each step certain portions of the DNA sample will not be recovered. Although optimizations of each step involved in PET construction could make incremental improvements, eventually, the PET method would have to be performed in automatic and miniaturized lab-on-a-chip systems to match the speed and efficiency of DNA sequencing machines. An important benefit of making PET constructs in a nanometer scale system is that this might allow PET analysis for smaller numbers of cells. Only with this nanoscale capability can PET analysis be applied to clinical samples that usually are not present in large quantities. The use of microfluidics technologies to manipulate tiny amounts of fluids using nanochannels (Whitesides 2006) would be necessary for the development of such miniaturized assays. Emulsion technologies could also be used to create "microreactors" of water droplets dispersed in oil for partitioning reactions (Griffiths and Tawfik 2006). This feature has been exploited in Roche 454 pyrosequencing and Applied Biosystems SOLiD ligation-based sequencing systems (Margulies et al. 2005) and should be applicable for reactions in PET protocols.

Increasing tag length of PET constructs is another aspect for improvement. The current PET preparation methods use tagging enzymes that are constrained by the restriction enzymes available (MmeI for 20 bp; EcoP15I for 27 bp). A theoretical analysis of sequences in the mouse genome suggests that 25-bp tags may have enough specificity to uniquely align with reference genome sequences (Faulkner et al. 2008). However, in reality, because of the complexity of nucleotide polymorphisms between individual genomes, repetitive and duplicated sequences in each genome (particularly mammalian genomes), and possible sequencing errors in individual sequence reads, less than half of tag (27 bp) or PET (27 bp/27 bp) sequences can be uniquely and perfectly mapped to the human reference genome sequences. With longer reads such as 36 bp and 75 bp, the unique and perfect tag mapping rate increases to 60%–70%, suggesting that longer tags could provide additional specificity for accurate mapping. To accommodate SNPs and sequencing errors, the general practice is to allow up to two mismatches for tag localization, which increases the mapping rate. Given that many tag sequences remain either unmapped or mapped to multiple locations, ideally, restriction enzymes that can cut longer tags or randomized enzymatic and nonenzymatic approaches should be developed to generate longer tag sizes in a PET construct. However, making tag sequences too long would lose the inherited efficiency of tag-based sequencing. It is our view that the ideal tag size for PET sequence mapping is 50–100 bp, which could offer an optimal balance of sequencing efficiency and mapping specificity.

With these PET improvements and continuing advances in sequencing technologies, we expect that PET-based methods will become the method of choice for many sequencing projects. Particularly, PET technology has great potential to make big contributions to the field of personal genomics. In the near future, more refined DNA-PET protocols would be combined with ultra-high-throughput sequencing technologies to give rise to a robust, cost-effective platform for individual personal human genome sequencing. In addition, the wide variety of PET applications for genome structure, transcriptome, and interactome characterizations will be useful in annotating the human genomes in great detail for functional and clinical implementations. With these new capacities, personal genome sequences combined with patient-specific transcriptomes and interactomes could become a practical reality and greatly benefit human healthcare and society.

In conclusion, the PET technology is a versatile method that can couple methods for asking biological questions with next-generation sequencing. With sequencing improving rapidly and increasing demands for sequencing to interrogate biological and clinical questions, the future of PET technologies is very bright.

Acknowledgments

The authors are supported by A*STAR of Singapore. In addition, M.J.F. is supported by A*STAR National Science Scholarships. Y.R and C.-L.W. are supported by NIH ENCODE grants R01HG003521-01, R01HG004456-01, and part of U54 HG004557-01.

References

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., et al. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252:** 1651–1656.

Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C., and Venter, J.C. 1992. Sequence identification of 2,375 human brain genes. *Nature* **355:** 632–634.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster. Science* **287:** 2185–2195.

Al-Dhaheri, M.H., Shah, Y.M., Basrur, V., Pind, S., and Rowan, B.G. 2006. Identification of novel proteins induced by estradiol, 4-hydroxytamoxifen and acolbifene in T47D breast cancer cells. *Steroids* **71:** 966–978.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

Bashir, A., Volik, S., Collins, C., Bafna, V., and Raphael, B.J. 2008. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Comput. Biol.* **4:** e1000051. doi: 10.1371/journal.pcbi.1000051.

Bourque, G., Leong, B., Vega, V.B., Chen, X., Lee, Y.L., Srinivasan, K.G., Chew, J.L., Ruan, Y., Wei, C.L., Ng, H.H., et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* **18:** 1752–1762.

Bovee, D., Zhou, Y., Haugen, E., Wu, Z., Hayden, H.S., Gillett, W., Tuzun, E., Cooper, G.M., Sampas, N., Phelps, K., et al. 2008. Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat. Genet.* **40:** 96–101.

Branco, M.R. and Pombo, A. 2006. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol.* **4:** e138. doi: 10.1371/journal.pbio.0040138.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18:** 630–634.

Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C., et al. 2008. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet.* **40:** 722–729.

Carninci, P. and Hayashizaki, Y. 1999. High-efficiency full-length cDNA cloning. *Methods Enzymol.* **303:** 19–44.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563.

Chen, J., Kim, Y.C., Jung, Y.C., Xuan, Z., Dworkin, G., Zhang, Y., Zhang, M.Q., and Wang, S.M. 2008a. Scanning the human genome at kilobase resolution. *Genome Res.* **18:** 751–762.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., et al. 2008b. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133:** 1106–1117.

Chiu, K.P., Wong, C.H., Chen, Q., Ariyaratne, P., Ooi, H.S., Wei, C.L., Sung, W.K., and Ruan, Y. 2006. PET-Tool: A software suite for comprehensive processing and managing of Paired-End diTag (PET) sequence data. *BMC Bioinformatics* **7:** 390. doi: 10.1186/1471-2105-7-390.

Collins, F.S. and Weissman, S.M. 1984. Directional cloning of DNA fragments at a large distance from an initial probe: A circularization method. *Proc. Natl. Acad. Sci.* **81:** 6812–6816.

Collins, F.S., Drumm, M.L., Cole, J.L., Lockwood, W.K., Vande Woude, G.F., and Iannuzzi, M.C. 1987. Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science* **235:** 1046–1049.

Cremer, T. and Cremer, C. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.* **2:** 292–301.

Cullen, K.E., Kladde, M.P., and Seyfred, M.A. 1993. Interaction between transcription regulatory regions of prolactin chromatin. *Science* **261:** 203–206.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. 2002. Capturing chromosome conformation. *Science* **295:** 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., et al. 2006. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16:** 1299–1309.

Dunn, J.J., McCorkle, S.R., Praissman, L.A., Hind, G., Van Der Lelie, D., Bahou, W.F., Gnatenko, D.V., and Krause, M.K. 2002. Genomic signature tags (GSTs): A system for profiling genomic DNA. *Genome Res.* **12:** 1756–1765.

Dunn, J.J., McCorkle, S.R., Everett, L., and Anderson, C.W. 2007. Paired-end genomic signature tags: A method for the functional analysis of genomes and epigenomes. *Genet. Eng. (N.Y.)* **28:** 159–173.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. 2008. Real-time DNA sequencing from single polymerase molecules. *Science* **323:** 133–138.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Euskirchen, G.M., Rozowsky, J.S., Wei, C.L., Lee, W.H., Zhang, Z.D., Hartman, S., Emanuelsson, O., Stolc, V., Weissman, S., Gerstein, M.B., et al. 2007. Mapping of transcription factor binding regions in mammalian cells by ChIP: Comparison of array- and sequencing-based technologies. *Genome Res.* **17:** 898–909.

Faulkner, G.J., Forrest, A.R., Chalk, A.M., Schroder, K., Hayashizaki, Y., Carninci, P., Hume, D.A., and Grimmond, S.M. 2008. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics* **91:** 281–288.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512.

Fraser, P. and Bickmore, W. 2007. Nuclear organization of the genome and the potential for gene regulation. *Nature* **447:** 413–417.

Griffiths, A.D. and Tawfik, D.S. 2006. Miniaturising the laboratory in emulsion droplets. *Trends Biotechnol.* **24:** 395–402.

Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J.W., et al. 2008. Single-molecule DNA sequencing of a viral genome. *Science* **320:** 106–109.

Hashimoto, S., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S., and Matsushima, K. 2004. 5′-End SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.* **22:** 1146–1149.

Holt, R.A. and Jones, S.J. 2008. The new paradigm of flow cell sequencing. *Genome Res.* **18:** 839–846.

Hong, G.F. 1981. A method for sequencing single-stranded cloned DNA in both directions. *Biosci. Rep.* **1:** 243–252.

Horike, S., Cai, S., Miyano, M., Cheng, J.F., and Kohwi-Shigematsu, T. 2005. Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. *Nat. Genet.* **37:** 31–40.

Hudson, M.E. and Snyder, M. 2006. High-throughput methods of regulatory element discovery. *Biotechniques* **41:** 673–681.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P., and Gingeras, T.R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296:** 916–919.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453:** 56–64.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. 2005. A high-resolution map of active promoters in the human genome. *Nature* **436:** 876–880.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420–426.

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:** 860–921.

Li, H., Ruan, J., and Durbin, R. 2008a. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18:** 1851–1858.

Li, R., Li, Y., Kristiansen, K., and Wang, J. 2008b. SOAP: Short oligonucleotide alignment program. *Bioinformatics* **24:** 713–714.

Lim, C.A., Yao, F., Wong, J.J., George, J., Xu, H., Chiu, K.P., Sung, W.K., Lipovich, L., Vega, V.B., Chen, J., et al. 2007. Genome-wide mapping of RELA(p65) binding identifies E2F1 as a transcriptional activator recruited by NF-κB upon TLR4 activation. *Mol. Cell* **27:** 622–635.

Lin, C.Y., Vega, V.B., Thomsen, J.S., Zhang, T., Kong, S.L., Xie, M., Chiu, K.P., Lipovich, L., Barnett, D.H., Stossi, F., et al. 2007. Whole-genome cartography of estrogen receptor αbinding sites. *PLoS Genet.* **3:** e87. doi: 10.1371/journal.pgen.0030087.

Ling, J.Q., Li, T., Hu, J.F., Vu, T.H., Chen, H.L., Qiu, X.W., Cherry, A.M., and Hoffman, A.R. 2006. CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. *Science* **312:** 269–272.

Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., et al. 2006. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38:** 431–440.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18:** 1509–1517.

Mastrangelo, I.A., Courey, A.J., Wall, J.S., Jackson, S.P., and Hough, P.V. 1991. DNA looping and Sp1 multimer links: A mechanism for transcriptional synergism and enhancement. *Proc. Natl. Acad. Sci.* **88:** 5670–5674.

Matsumura, H., Reich, S., Ito, A., Saitoh, H., Kamoun, S., Winter, P., Kahl, G., Reuter, M., Kruger, D.H., and Terauchi, R. 2003. Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proc. Natl. Acad. Sci.* **100:** 15718–15723.

Mauro, M.J., O'Dwyer, M., Heinrich, M.C., and Druker, B.J. 2002. STI571: A paradigm of new agents for cancer therapeutics. *J. Clin. Oncol.* **20:** 325–334.

Metzker, M.L. 2005. Emerging technologies in DNA sequencing. *Genome Res.* **15:** 1767–1776.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., et al. 2007. Genome-wide

maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553–560.
Milner, R.J. and Sutcliffe, J.G. 1983. Gene expression in rat brain. *Nucleic Acids Res.* **11:** 5497–5520.
Mitelman, F., Johansson, B., and Mertens, F. 2007. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* **7:** 233–245.
Morgan, R.D., Bhatia, T.K., Lovasco, L., and Davis, T.B. 2008. MmeI: A minimal Type II restriction-modification system that only modifies one DNA strand for host protection. *Nucleic Acids Res.* **36:** 6558–6570. doi: 10.1093/nar/gkn711.
Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., and Marra, M. 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45:** 81–94.
Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5:** 621–628.
Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., et al. 2000. A whole-genome assembly of *Drosophila*. *Science* **287:** 2196–2204.
Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.
Ng, P., Wei, C.L., Sung, W.K., Chiu, K.P., Lipovich, L., Ang, C.C., Gupta, S., Shahab, A., Ridwan, A., Wong, C.H., et al. 2005. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* **2:** 105–111.
Ng, P., Tan, J.J., Ooi, H.S., Lee, Y.L., Chiu, K.P., Fullwood, M.J., Srinivasan, K.G., Perbost, C., Du, L., Sung, W.K., et al. 2006. Multiplex sequencing of paired-end ditags (MS-PET): A strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Res.* **34:** e84. doi: 10.1093/nar/gkl444.
Ng, P., Wei, C.L., and Ruan, Y. 2007. Paired-end ditagging for transcriptome and genome analysis. *Curr. Protoc. Mol. Biol.* **21:** 21.12.1–21.12.42.
Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W., et al. 2004. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat. Genet.* **36:** 1065–1071.
Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20:** 207–211.
Pop, M. and Salzberg, S.L. 2008. Bioinformatics challenges of new sequencing technology. *Trends Genet.* **24:** 142–149.
Putney, S.D., Herlihy, W.C., and Schimmel, P. 1983. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302:** 718–721.
Raghavendra, N.K. and Rao, D.N. 2005. Exogenous AdoMet and its analogue sinefungin differentially influence DNA cleavage by R.EcoP15I—Usefulness in SAGE. *Biochem. Biophys. Res. Commun.* **334:** 803–811.
Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306–2309.
Ruan, Y., Ooi, H.S., Choo, S.W., Chiu, K.P., Zhao, X.D., Srinivasan, K.G., Yao, F., Choo, C.Y., Liu, J., Ariyaratne, P., et al. 2007. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res.* **17:** 828–838.
Rubin, G.M. and Lewis, E.B. 2000. A brief history of *Drosophila*'s contributions to genome research. *Science* **287:** 2216–2218.
Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W., and Velculescu, V.E. 2002. Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **20:** 508–512.
Schuster, S.C. 2008. Next-generation sequencing transforms today's biology. *Nat. Methods* **5:** 16–18.
Shastry, B.S. 2007. SNPs in disease gene mapping, medicinal drug development and evolution. *J. Hum. Genet.* **52:** 871–880.
Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309:** 1728–1732.
Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci.* **100:** 15776–15781.
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:** 1348–1354.
Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321:** 956–960.
Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol. Cell* **10:** 1453–1465.
Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al. 2005. Fine-scale structural variation of the human genome. *Nat. Genet.* **37:** 727–732.
van der Hage, J.A., van den Broek, L.J., Legrand, C., Clahsen, P.C., Bosch, C.J., Robanus-Maandag, E.C., van de Velde, C.J., and van de Vijver, M.J. 2004. Overexpression of P70 S6 kinase protein is associated with increased risk of locoregional recurrence in node-negative premenopausal early breast cancer patients. *Br. J. Cancer* **90:** 1543–1550.
Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. 1995. Serial analysis of gene expression. *Science* **270:** 484–487.
Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett Jr., D.E., Hieter, P., Vogelstein, B., and Kinzler, K.W. 1997. Characterization of the yeast transcriptome. *Cell* **88:** 243–251.
Venter, J.C., Smith, H.O., and Hood, L. 1996. A new strategy for genome sequencing. *Nature* **381:** 364–366.
Venter, J.C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O., and Hunkapiller, M. 1998. Shotgun sequencing of the human genome. *Science* **280:** 1540–1542.
Wang, T.L., Maierhofer, C., Speicher, M.R., Lengauer, C., Vogelstein, B., Kinzler, K.W., and Velculescu, V.E. 2002. Digital karyotyping. *Proc. Natl. Acad. Sci.* **99:** 16156–16161.
Weber, J.L. and Myers, E.W. 1997. Human whole-genome shotgun sequencing. *Genome Res.* **7:** 401–409.
Wei, C.L., Ng, P., Chiu, K.P., Wong, C.H., Ang, C.C., Lipovich, L., Liu, E.T., and Ruan, Y. 2004. 5′ Long serial analysis of gene expression (LongSAGE) and 3′ LongSAGE for transcriptome characterization and genome annotation. *Proc. Natl. Acad. Sci.* **101:** 11701–11706.
Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z., et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124:** 207–219.
Whitesides, G.M. 2006. The origins and the future of microfluidics. *Nature* **442:** 368–373.
Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bahler, J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453:** 1239–1243.
Wold, B. and Myers, R.M. 2008. Sequence census methods for functional genomics. *Nat. Methods* **5:** 19–21.
Woodcock, C.L. 2006. Chromatin architecture. *Curr. Opin. Struct. Biol.* **16:** 213–220.
Wurtele, H. and Chartrand, P. 2006. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Res.* **14:** 477–495.
Yoshimura, S.H., Maruyama, H., Ishikawa, F., Ohki, R., and Takeyasu, K. 2004. Molecular mechanisms of DNA end-loop formation by TRF2. *Genes Cells* **9:** 205–218.
Zeller, K.I., Zhao, X., Lee, C.W., Chiu, K.P., Yao, F., Yustein, J.T., Ooi, H.S., Orlov, Y.L., Shahab, A., Yong, H.C., et al. 2006. Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proc. Natl. Acad. Sci.* **103:** 17834–17839.
Zerbino, D.R. and Birney, E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18:** 821–829.
Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U., et al. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38:** 1341–1347.
Zhao, X.D., Xu, H., Chew, J.L., Liu, J., Chiu, K.P., Choo, A., Orlov, Y.L., Sung, K.W., Shahab, A., Kuznetsov, V.A., et al. 2007. Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* **1:** 286–298.