UNIVERSITY OF ZAGREB
FACULTY OF SCIENCE
DIVISION OF BIOLOGY

EXPERIMENTAL METHODS IN FUNCTIONAL GENOMICS

SEMINAR PAPER

Ivan Mikičić
Undergraduate Study of Molecular Biology
Mentor: prof. dr. sc. Kristian Vlahoviček

Zagreb, 2015

TABLE OF CONTENTS

1. INTRODUCTION.....3
2. GENOMICS – DNA SEQUENCING.....3
3. TRANSCRIPTOME PROFILING.....7
4. EPIGENOMICS.....10
    - 4.1.PROTEIN-DNA INTERACTIONS AND COVALENT HISTONE MODIFICATIONS.....10
    - 4.2.CHROMATIN INTERACTIONS.....13
5. ENCODE.....15
6. CONCLUSION.....15
7. LITERATURE.....16
8. SUMMARY.....17

1. INTRODUCTION

The end of the 20th century marks the development of techniques that enable rapid data production in life sciences and consequently development of omics approaches. Omics refers to collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms. One of the omics disciplines is functional genomics, the aim of which is to understand the complex relationship between genotype and phenotype on a global (genome-wide) scale. Functional genomics studies investigate a range of processes such as transcription, translation and epigenetic regulation, as opposed to the static aspects of genomic information, such as DNA sequence or structure, that fall under the domain of genomics. To explore the cappacities of functional genomics approaches, new methods had to be invented with the ability to assess function on a genomic level that are statistically robust and high-throughput at the same time. The aim of this review is to give an overview of the most common laboratory techniques used in functional genomics, starting on the DNA and RNA level and finishing with various methods that aim at discovering interactions, such as protein-DNA and chromatin interactions.

2. GENOMICS – DNA SEQUENCING

DNA sequencing is the process of determining the order of bases in a length of DNA. Whole genome sequencing focuses on sequencing all the DNA in an organism's genome. During the last few decades an enormous progress has been made in the development of new sequencing technologies, with the aim of making sequencing faster and cheaper. Although DNA sequencing itself does not in the real sense fall within the scope of functional genomics, sequenced genomes are a prerequisite for any further functional investigation and therefore DNA sequencing deserves to be covered in this review. The era of DNA sequencing started with the Sanger method, which is based on the selective incorporation of flourecently labeled chain-terminating dideoxynucleotides (ddNTPs) by DNA polymerase during a PCR reaction. In the resulting mixture of DNA molecules of various sizes, the different fluorescent dyes mark different ddNTPs at the end of each DNA molecule. The DNA sequence is determined by high-resolution electrophoretic separation of DNA fragments and assessment of their length and ddNTP identity. Modifications of this method dominated the sequencing industry

for almost two decades and enabled sequencing of the first finished-grade human genome sequence.

The Sanger method led to a number of monumental accomplishments but it was also slow, expensive and labor-intensive. Because of this, newer methods, referred to as the next-generation sequencing (NGS), were developed. The advancements in NGS are reflected in increasing output per run, read lengths and accuracy of base-calling as well as decreasing the costs of sequencing. Second- and third-generation sequencers offered the ability to produce an enormous volume of data cheaply. As a result, the rate limiting step in answering to biological questions today is the sequence analysis, not the sequence production. One consequence of the sequencing revolution is the revitalization of bioinformatics, predominantly in efforts aimed at data analysis and interpretation.

Although each platform is different in its specifics, all NGS devices share certain attributes. Firstly, the preparatory steps are fewer and simpler than for Sanger sequencing. Instead of beginning with bacterial cloning and DNA isolation, NGS experiment begins with the creation of template library. Templates are formed by ligating platform-specific synthetic DNA molecules (adapters) onto the ends of the fragment population to be sequenced. Secondly, the templates need to be immobilized on a solid surface for amplification by a polymerase-mediated reaction. Lastly, the sequencing reactions are carried out in a series of repeating steps and the products are detected automatically.

Sequencing technologies include a number of methods that are grouped broadly as template preparation, sequencing, imaging and data analysis. Current methods for template preparation generally involve randomly breaking genomic DNA into smaller sizes from which the templates are created. The immobilization of spatially separated templates allows thousands to billions of sequencing reactions to be performed simultaneously. Templates can be clonally amplified or single-molecule. Latter approach does not require PCR, which creates mutations that can be mistaken for single nucleotide polymorphisms (SNPs). Moreover, AT- and GC-rich sequences show amplification bias so the quantitative approaches, such as RNA-seq, perform better with single-molecule templates.

In the past couple of years, Illumina sequencers have gained most of the market share due to the best balance between read lengths, error rates and cost as well as rapid improvements in sequencing chemistry and availability of more extensive bioinformatics tools. Because of this, I believe that the Illumina platform deserves to be described in this review as an example of distinct sequencing strategy. During template preparation, Illumina uses the so called solid-phase amplification procedure, which is composed of two steps: initial

priming and extending of the single-stranded, single-molecure template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. In the sequencing and imaging step Illumina uses the cyclic reversible termination technology, which comprises nucleotide incorporation, fluorescence imaging and cleavage. DNA polymerase bound to the primed template adds just one fluorescently modified nucleotide, which represents complement of the template base. The remaining unincorporated nucleotides are then washed away and the imaging is performed to determine the identity of the incorporated nucleotide. This is followed by a cleavage step, which removes the terminating/inhibiting group and the fluorescent dye.

After NGS reads have been generated, they are aligned to a known reference sequence or assembled *de novo*. The choice of the strategy depends on biological application, cost, effort and time considerations. Alignment to a reference sequence is cheaper, faster and less time-consuming but it is also limited by factors such as placing reads within repetitive regions and nonexistent corresponding regions in the reference sequence. These factors can be partially overcome by using mate-pair reads, increasing the reading coverage (which results in increasing contig length and decreasing ambiguities in an assembly) and combining different NGS platform reads.

Genome assembly is followed by genome annotation, the process of identifying genes and their intron-exon structures. Although the sequencing has become easier with the introduction of NGS technologies, genome annotation has become more challenging. There are several reasons for this: shorter read lengths of NGS technologies, exotic nature of many recently sequenced genomes and the need to update already annotated genomes with new data.

The first step towards the successful annotation of a genome is to determine whether its assembly is ready for annotation. There are several summary statistics that can be used to assess the completeness and contiguity of a genome assembly – N50, average gap size of a scaffold and average number of gaps per scaffold. The most commonly used summary statistics is N50. A contig N50 is calculated by summarizing the lengths of contigs starting from the longest one, until the sum equals one half of all contigs in that assembly. The contig N50 is the length of the shortest contig in this list. The scaffold N50 is calculated in the same manner but uses scaffolds instead of contigs. Although there are no strict rules, an assembly with an N50 scaffold length that is gene-sized is a decent target for annotation because this means that around 50% of the genes will be contained on a single scaffold. If an assembly is

incomplete or its N50 scaffold length is too short, additional shotgun sequencing is recommendable.

Generally, genome-wide structural annotation can be divided into two phases, computational and annotation phase. The so called computational phase includes repeat identification and masking followed by alignment with already annotated sequences, such as ESTs, RNA-seq data and protein and DNA sequences from databases. Repeats complicate genome annotation and should therefore be masked. The term masking implies transforming every nucleotide identified as a repeat to an 'N' or, for soft masking, lower case a, t, c or g. The importance of repeat masking cannot be overemphasized: unmasked repeats can produce a number of false BLAST alignments. On top of that, many transposon ORFs look like true host genes and can be added as additional exons to gene predictions. Gene prediction can be *ab initio*, using mathematical models rather than external evidence, or evidence-driven. Most of today's gene prediction tools use combination of *ab initio* and evidence-driven methods.

The annotation phase includes obtaining a synthesis of alignment-based evidence with *ab initio* gene predictions to obtain a final set of gene annotations. Traditionally, this was done manually. This is the most precise but also the most time-consuming way of gene prediction so more and more laboratories are using automated annotation pipelines. This implies running a number of different gene finders and using a chooser algorithm (also known as combiner) to select the prediction whose intron-exon structure best represents the consensus of the models.

The last step of the annotation is the visualization of the annotation data. There are several commonly used output formats for describing annotations, including GenBank, GFF3, GTF and EMBL. FASTA format shouldn't be used as the output format for annotations because it is insufficiently informative and enables only a small subset of possible downstream analyses. There is one more essential part of the annotation process, and that is the quality control. Once a genome is annotated, it doesn't mean that all the annotations are correct – even the most precise gene predictors and pipelines manage to accomplish the accuracy of around 80%. Flawed annotations are dangerous because they are transferred to future annotations and can lead to false conclusions when used in research.

Genome annotations find their use in medicine (discovery of clinically significant variants, treatment personalization), in seq-based methods (ChIP-seq, DNase-seq, methyl-seq,

RNA-seq), metagenomics and are a prerequisite for all downstream investigations, such as transcriptomics, proteomics and epigenomics.

Despite their limitations, advances in next-generation sequencing have been so rapid that they have ceased to amaze. New single-molecule methods, particularly nanopore sequencing, promise to significantly expand the realm of possibilities by greatly decreasing costs. Recently, a startup called Oxford Nanopore announced the launch of a new third generation single-molecule sequencing platform. By moving single-pore sequencing technologies to market, Oxford Nanopore promised read lengths orders of magnitude longer than existing technologies, together with low per-base costs, and a tiny futuristic USB-powered sequencer. The hand-held $1000 sequencer, with a simple library preparation process, promised to democratize sequencing, making it affordable to a larger community and perhaps even to citizen-scientists.

3. TRANSCRIPTOME PROFILING

The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition. Various technologies have been developed to deduce and quantify the transcriptome. Traditionally wide-used set of methods can be characterized as hybridization-based approaches, with the most commonly used application being microarrays. Microarray is a hybridization of a nucleic acid sample (target) to a very large set of oligonucleotide probes, which are attached to a solid support, to determine sequence or to detect variations in a gene sequence or expression or for gene mapping. Hybridization-based approaches are high-throughput and relatively inexpensive. On the other side, they are based on previously discovered sequences, have high background levels, limited dynamic range of detection, complicated normalization, low resolution, require high quantities of RNA and are unsuitable for detailed transcriptomics of large genomes.

Recently, the development of novel high-throughput DNA sequencing methods has provided a new method for both mapping and quantifing transcriptomes, named RNA-sequencing (or RNA-seq). This method includes reversely transcribing a population of RNA molecules into cDNA with adaptors on one or both ends, which is followed by single-end or paired-end sequencing. The obtained reads are aligned to a reference genome or assembled *de novo*.

RNA-seq has several advantages compared to microarrays – it is not limited to detecting transcripts that correspond to a known genomic sequence, it enables discovery of exon connectivity (long RNA-seq reads exceed the borders of a single exon) and SNPs (due to high resolution), it has low background levels, high dynamic range with no upper detection limit and it requires less RNA material. However, there are several challenges that need to be taken into consideration when performing RNA-seq experiments. RNA fragmentation introduces bias into detection as the body of a sequence is sequenced more precisely than the ends. cDNA fragmentation introduces a different kind of bias because the 3'-end is sequenced more precisely. Amplification of fragments leads to bias in quantification; some regions are amplified more easily so the quantity ratios of initial RNA sequences are changed during amplification. Bioinformatic challenges of RNA-seq manifest in mapping alternative and trans-splicing and assessing repeats which need to be mapped to more than one locus in the genome. The latter problem can be solved by higher read lengths to include the flanking unique sequences.

It is important to mention sequence coverage, or the percentage of the transcripts surveyed. Greater coverage requires more sequencing depth but enables more percise applications, such are detection of rare transcripts or variants. Generally, the more complex the transcriptome, the more sequencing depth is required for adequate coverage. Despite the challenges described above, the RNA-seq has enabled insight into some previously unexplored areas, such as unknown transcribed regions (especially 3'- and 5'-UTRs), identification of isoforms, precise 3'- and 5'-end mapping and absolute quantification of gene expression.

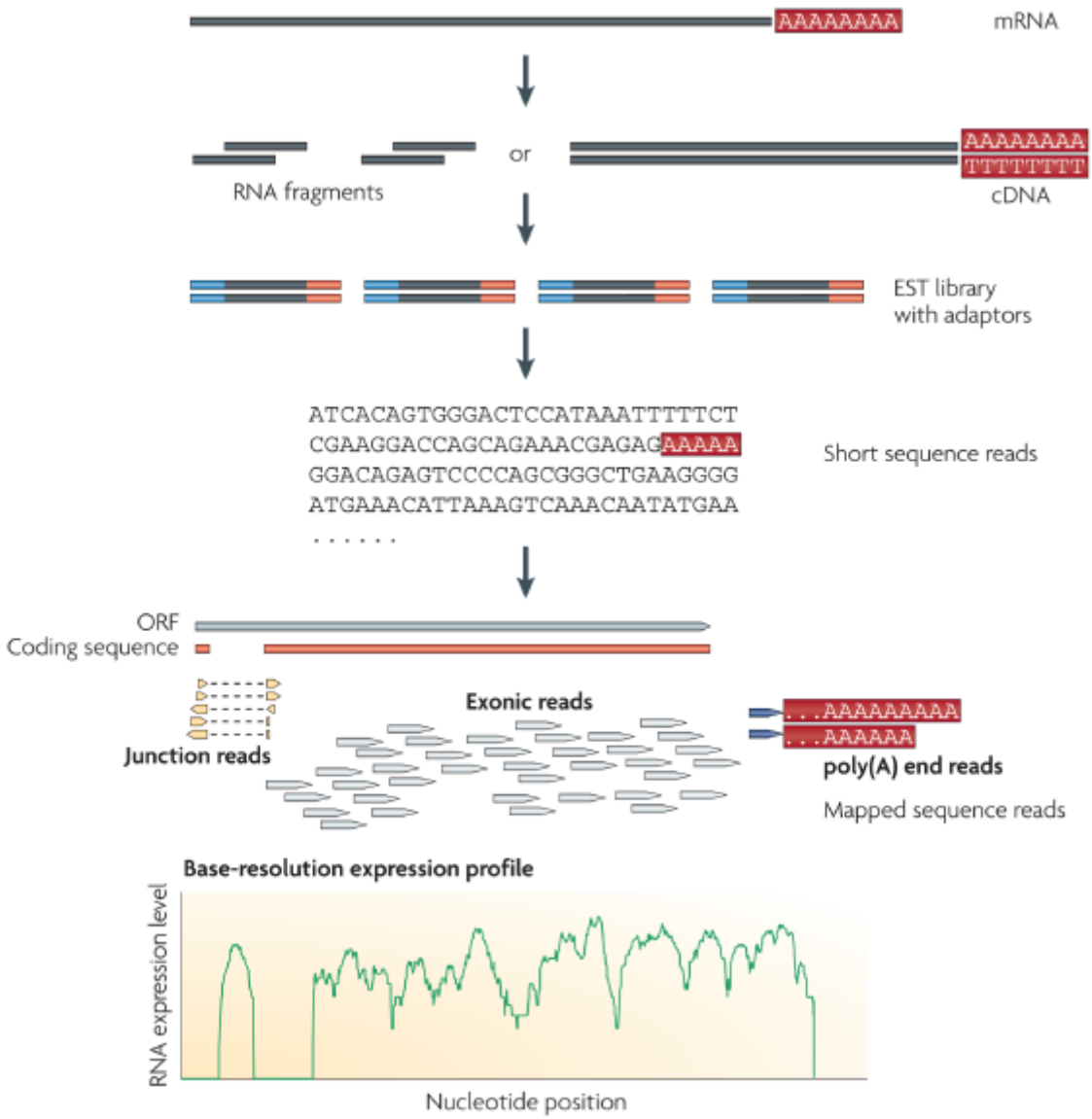

Figure 1. A typical RNA-seq experiment (courtesy of Wang et al. 2009). Long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom; a yeast ORF with one intron is shown.

4. EPIGENOMICS

4.1. PROTEIN-DNA INTERACTIONS AND COVALENT HISTONE MODIFICATIONS

Epigenomics is the study of heritable changes other than those in the DNA sequence on a genome-wide level and encompasses two major modifications of DNA or chromatin: DNA methylation, the covalent modification of cytosine, and post-translational modification of histones including methylation, acetylation, phosphorylation and sumoylation.

Determining how proteins interact with DNA to regulate gene expression is essential for fully understanding many biological processes and disease states. This epigenetic information is complementary to genotype and expression analysis. DNA-binding proteins play important roles in many cellular pocesses, including transcription, splicing, replication, DNA repair and condensation. Locations of bound factors and histone modifications cannot be precisely predicted from the DNA sequence alone so functional methods are necessary for detection of these characteristics. Chromatin immunoprecipitation coupled with short-tag sequencing (ChIP-seq) has become the standard method for detection of protein-DNA interactions and histone modifications. It has largely replaced previously used, hybridization-based approach called ChIP-chip, which has become somewhat redundant with development of sequencing technologies.

There are two versions of the most common ChIP-seq protocol, one intended for DNA-binding proteins and the other for histone modifications. They include crosslinking, fragmentation, immunoprecipitation and purification, sequencing and detection. A more detailed description can be found in the Figure 2.

The success of a ChIP-seq experiment depends on the development of highly specific antibodies and requires a high number of cells (there are technologies coping with this problem, such as Nano-ChIP-seq and LinDA). Moreover, fragments derived by sonication are typically ~200 bases in length while each protein typically binds only 6-20 bases. There are special versions of ChIP-seq developed to cope with the problem of resolution, eg. exonuclease from phage lambda is used to remove excess sequences flanking the binding site. The question of multiple proteins binding to the same site (they could be binding simultaneously or at different times) is solved by sequential ChIP-seq.

Pipelines for ChIP-seq analysis are computational protocols that serve for detection and visualization of ChIP-seq experiment results. Regions of protein-DNA interaction or

histone modification are identified as DNA sites that contain enriched signals, that is, where more sequences are aligned than would be expected by chance. Several characteristic challenges need to be taken into account during analysis of ChIP-seq data. Mappability is the uniqueness of a stretch of DNA sequence compared with a whole-genome sequence. If the detected region is unique, it is easily mappable and vice versa. The problem of mappability can partially be solved by using paired-end reads. Another specificity of ChIP-seq data is the existance of three peak types – point source, broad source and mixed source. Each type requires a distinct strategy for detection. The need to compare data obtained in various experiments further complicates the ChIP-seq analysis but it serves as a way to extract valuable biological information, such as to detect differences in protein-DNA interactions and histone modifications in different cell types or environmental conditions.

There are also a few analytical challenges. For example, signal strength depends on the strength of the interaction, which in turn depends on the genotype. To acquire the signal mean, cell from various individuals should be assayed. Moreover, the signals do not always represent direct but also indirect binding due to the proximity of the assayed protein and a non-binding sequence during the crosslinking step. This difference cannot be distinguished using just ChIP-seq.

Most transcription factors cannot stably interact with binding sites if DNA is nucleosomal. This means that nucleosome-depleted, open chromatin regions are necessary for stable binding to occur. Detecting open chromatin complements ChIP-seq and two methods, DNase-seq and FAIRE-seq, have been developed to detect open chromatin directly.

DNase-seq is based on the property of DNase to digest only open chromatin. It has been shown that open chromatin mainly corresponds to regulatory regions, like promoters, enhancers, silencers and insulators so this method can implicate functionality of the regions enriched in DNase digestion. FAIRE-seq is based on crosslinking, sonication and phenol-chloroform extraction, after which the nucleosome-depleted fraction is segregated to the aqueous phase. These methods are not based on protein-DNA interaction, but they help to identify regions with potential protein-binding ability that were not discovered with ChIP-seq due to the lack of specific antibodies.

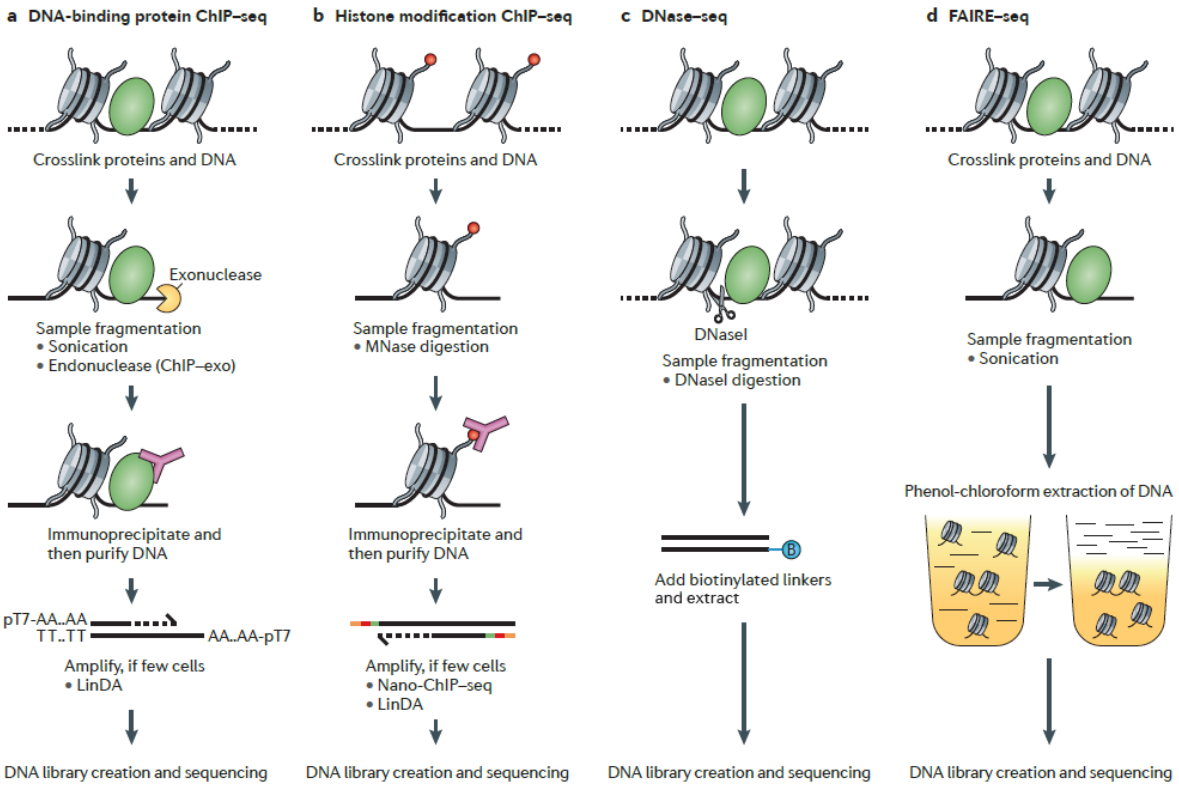

Figure 2. Comparison of experimental protocols (courtesy of Furey, 2012). Experiments to detect different aspects of DNA-binding proteins share many of the same steps; simplified schematics of the main steps are shown. **a** Chromatin immunoprecipitation followed by sequencing (ChIP–seq) for DNA-binding proteins such as transcription factors. Recent variations on the standard protocol include using endonuclease digestion instead of sonication (ChIP–exo) to increase the resolution of binding-site detection and to eliminate contaminating DNA, and DNA amplification after ChIP for samples with limited cells. **b** ChIP–seq for histone modifications uses micrococcal nuclease (MNase) digestion to fragment DNA and can also now be run on low-quantity samples when combined with the additional post-ChIP amplification. **c** DNase–seq relies on digestion by the DNaseI nuclease to identify regions of nucleosome-depleted open chromatin where there are binding sites for all types of factors, but it cannot identify what specific factors are bound. **d** Formaldehyde-assisted identification of regulatory elements (FAIRE–seq) similarly identifies nucleosome-depleted regions by extracting fragmented DNA that is not crosslinked to nucleosomes. LinDA, single-tube linear DNA amplification; T7, T7 phage RNA polymerase.

4.2. CHROMATIN STRUCTURE

Methods for chromatin interactions mapping enable comprehension of regulatory programs and other high-complexity processes in cells that cannot be understood by using ChIP-seq, DNase-seq or FAIRE-seq. This includes approaches based on the chromatin conformation capture (commonly known as 3C) method – 3C, 4C, 5C, Hi-C and ChIA-PET. The aim of these approaches is to determine the frequency with which any pair of loci in the genome is in close enough physical proximity to become crosslinked and the principal steps in 3C-based experiments are to crosslink genomic regions that are in close proximity, digest DNA using restriction enzymes, which results in pairs of crosslinked DNA that came from distinct genomic locations and finally to identify these paires, for example via paired-end sequencing.

3C requires PCR primers for regions of interest (low-throughput), 5C simultaneously uses thousands of primers to detect millions of interactions (all-by-all, high-throughput) and Hi-C does not depend on primers but uses biotinylated residues (genome-wide method). ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) is used to identify interactions that are associated with specific proteins by using specific antibodies. It provides genome-wide high-resolution data for a given DNA-binding protein. The problem of all 3C based methods is that the resolution is limited by the frequency and distribution of restriction sites.

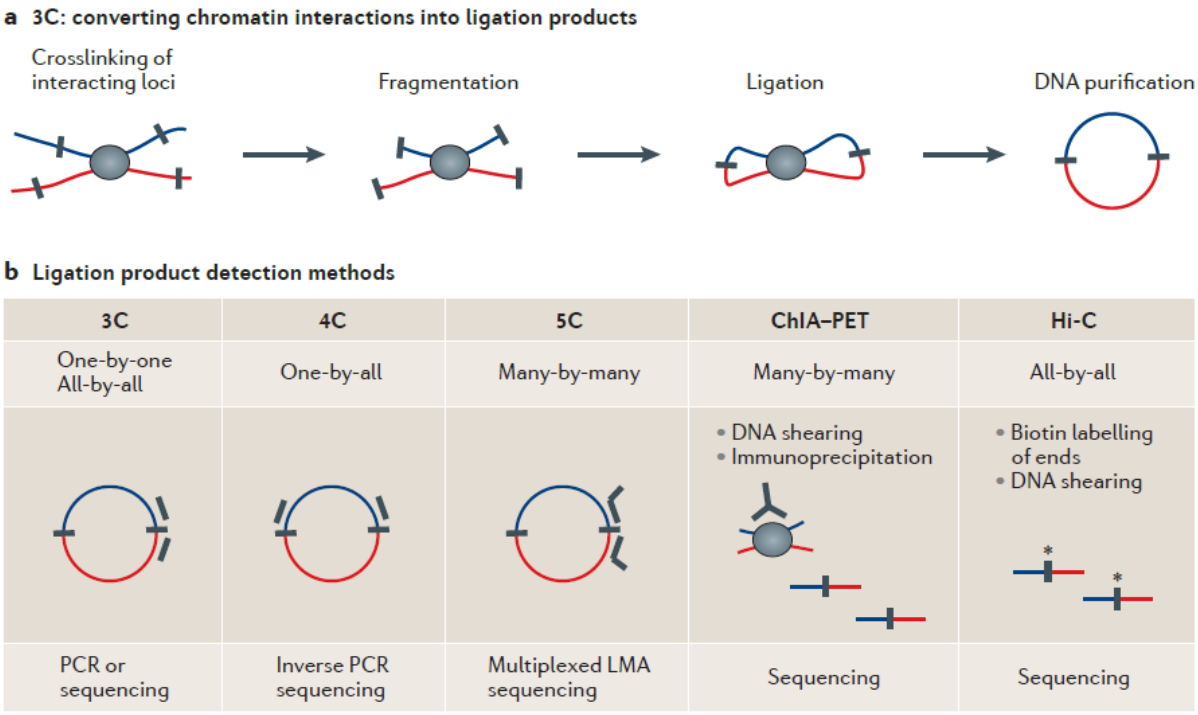

Figure 3. 3C-based technologies (courtesy of Dekker et al. 2013). In chromosome conformation capture (3C)-based methods (panel **a** of the figure), cells are crosslinked with formaldehyde to link chromatin segments covalently that are in close spatial proximity. Next, chromatin is fragmented by restriction digestion or sonication. Crosslinked fragments are then ligated to form unique hybrid DNA molecules. Finally, the DNA is purified and analysed. The different 3C-based methods differ only in the way that hybrid DNA molecules, each corresponding to an interaction event of a pair of loci, are detected and quantified (panel **b** of the figure). In classical 3C experiments, single ligation products are detected by PCR one at the time using locus-specific primers. Given that 3C can be laborious, most 3C analyses typically cover only tens to several hundreds of kilobases. 4C (also known as 'circular 3C' or '3C-on-chip') uses inverse PCR to generate genome-wide interaction profiles for single loci. 5C combines 3C with hybrid capture approaches to identify up to millions of interactions in parallel between two large sets of loci: for example, between a set of promoters and a set of distal regulatory elements. 4C approaches are genome-wide but are anchored on a single locus. 5C analyses typically involve two sets of hundreds to thousands of restriction fragments to interrogate up to millions of long-range interactions that can cover up to tens of megabases and that can be contiguous or scattered among loci of interest throughout the genome. The Hi-C method was the first unbiased and genome-wide adaptation of 3C and includes a unique step in which, after restriction digestion, the staggered DNA ends are filled in with biotinylated nucleotides (as shown by the asterisks). This facilitates selective

purification of ligation junctions that are then directly sequenced. Hi-C provides a true all-by-all genome-wide interaction map, but the resolution of this map depends on the depth of sequencing. Finally, various approaches combine 3C with chromatin immunoprecipitation to enrich for chromatin interactions between loci bound by specific proteins of interest. For instance, ChIA–PET method allows for genome-wide analysis of long-range interactions between sites bound by a protein of interest. Because ChIA–PET data represent a selected subset of interactions that occur in the genome, the three analysis approaches described in this article cannot directly be applied to this data type. LMA, ligation-mediated amplification.

5. THE ENCODE PROJECT

ENCODE or the Encyclopedia of DNA elements is one of the biggest projects in life sciences, that emerged as a follow-up to the Human Genome Project. The aim of the project was to identify all regions of transcription, transcription factor association, chromatin structure and histone modification in the human genome or, in other words, to discover all functional elements encoded in the human genome. Functional element in genomic cotext is defined by the ENCODE project as a discrete genome segment that encodes a defined product or displays a biochemical signaure (protein binding, special chromatin structure etc.). ENCODE used a number of methods described in this review, such as RNA-seq, ChIP-seq, DNase-seq, FAIRE-seq, 5C and ChIA-PET.

Major discoveries of the ENCODE project include the following: 80% of the human genome is functional, promoter functionality can explain most of the variation in RNA expression, non-coding regions account for more functional elements than the coding regions, medically relevant SNPs are enriched within non-coding functional elements and there are 4 major groups of functional elements: sites coding for RNA molecules, regions enriched for histone modifications, regions of open chromatin and sites of transcription factor binding.

6. CONCLUSION

The NGS sequencing has, due to its cost effectiveness and its many different uses, emerged as the dominant genomics technology. The new sequencers have provided genome-scale sequencing capacity to individual laboratories in addition to larger genome centers, thus

making sequencing widely available. The short read structure of next-generation sequencers provides potential problems for sequence assembly particularly in areas associated with sequence repeats. The short read length also necessitates the development of paired-end sequencing approaches for improved mapping efficiency. The potential of NGS technology is noticeable in a wide range of aplications, from DNA-sequencing and RNA-sequencing to methods for assaying protein-DNA (ChIP-seq, DNase-seq, FAIRE-seq) and chromatin interactions (3C-based methods). Just how important functional genomics has become is reflected in one of the biggest life science projects of the 21st century, the ENCODE project. Although the next generation sequencers are already being widely used, there are other sequencing methods, such as nanopore sequencing, whose scalability is being explored to decrease the sequencing cost and enhance throughput even further.

7. LITERATURE

1.Dekker, J., Marti-Renom, M. A., Mirny, L. A. Exploring the three dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews Genetics, 2013.

2. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature, 2012.

3. Furey, T. S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nature Reviews Genetics, 2012.

4. Mardis, E. R. A decade's perspective on DNA sequencing technology. Nature, 2011.

5. Metzker, M. L. Sequencing technologies – the next generation. Nature Reviews Genetics, 2010.

6. Mikheyev, A. S., Tin M. M. Y. A first look at the Oxford Nanopore MinION sequencer. Molecular Ecology Resources, 2014.

7. Morozova, O., Marra, M. A. Applications of next-generation sequencing technologies in functional genomics. Genomics, 2008.

8. Ozsolak F., Milos, P. M. RNA sequencing: advances, challenges and opportunities. Nature Reviews Genetics, 2011.

9. van Steensel, B., Dekker, J. Genomics tools for unraveling chromosome architecture. Nature Biotechnology, 2010.

10. Wang, Z., Gerstein, M., Snyder, M. RNA-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 2010.

11. Yandell, M., Ence, D. A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics, 2012.

8. SUMMARY

Over the past few years, massively parallel DNA sequencing platforms have become widely available, dramatically reducing the cost of DNA sequencing. The fast and low-cost sequencing approaches not only change the landscape of genome sequencing projects but also usher in new opportunities for sequencing in various applications, among other in functional genomics. The aim of functional genomics studies is to understand the complex relationship between genotype and phenotype on a genome-wide scale. Studies investigate a range of processes such as transcription, translation and epigenetic regulation, in an attempt to answer relevant biological questions. Newly developed techniques, such as RNA-seq, ChIP-seq and 3C-based methods found their use in some of the most massive projects, such as the ENCODE project. The aim of this review was to introduce some of the most common techniques in functional genomics and to explain their possibilities, advances, applications and challenges.