Genomics 101 (2013) 86–93

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Review

Molecular methods for genotyping complex copy number polymorphisms

Stuart Cantsilieris [a,b], Paul N. Baird [a,1], Stefan J. White [b,*,1]

[a] Centre for Eye Research Australia, University of Melbourne, Royal Victorian Eye and Ear Hospital, East Melbourne, Victoria, Australia
[b] Centre for Reproduction and Development, Monash Institute of Medical Research, Monash University, Melbourne, Victoria, Australia

ARTICLE INFO

Article history:
Received 20 July 2012
Accepted 24 October 2012
Available online 30 October 2012

Keywords:
Copy number variation
Next generation sequencing
Paralogue ratio test
Multiplex ligation-dependent probe amplification
Quantitative PCR
Southern blotting
Multiplex amplifiable probe hybridization

ABSTRACT

Genome structural variation shows remarkable complexity with respect to copy number, sequence content and distribution. While the discovery of copy number polymorphisms (CNP) has increased exponentially in recent years, the transition from discovery to genotyping has proved challenging, particularly for CNPs embedded in complex regions of the genome. CNPs that are collectively common in the population and possess a dynamic range of copy numbers have proved the most difficult to genotype in association studies. This is in some part due to technical limitations of genotyping assays and the sequence properties of the genomic region being analyzed. Here we describe in detail the basis of a number of molecular techniques used to genotype complex CNPs, compare and contrast these approaches for determination of multi-allelic copy number, and discuss the potential application of these techniques in genetic studies.

Contents

1. Introduction

The human genome contains several levels of genetic variation, from single base changes to those affecting entire chromosomes. Copy number variants (CNVs) now operationally defined as deletions

* Corresponding author at: Centre for Reproduction and Development, Monash Institute of Medical Research, 27-31 Wright Street, Clayton, 3168, Victoria, Australia. Fax: +61 3 9594 7114.
E-mail address: stefan.white@monash.edu (S.J. White).
[1] Are to be recognized as joint senior authors.

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2012.10.004

and duplications >50 bp, can be rare (<1%) or common (>5%), large (>1Mp) or small (<500 bp), and di-allelic (0–3) or multi-allelic (>3 diploid copy numbers) [1]. CNVs that segregate at appreciable frequency in the population are termed copy number polymorphisms (CNPs), and those that show a dynamic range of diploid copy number (multi-allelic) will be the focus of this review. Multi-allelic CNPs are attractive candidates for disease association studies for several reasons. Many of these CNPs contain genes, and these genes appear to be over-represented in pathways associated with immunity and interaction with the environment [2]. Furthermore, there is evidence of stratification in human populations, indicating that such regions are positively selected and are of clinical relevance [3]. Finally, gene duplication followed by adaptive evolution can facilitate new gene function, resulting in changes of phenotype [4].

Several techniques have been described to measure CNVs in the human genome, however, with such diverse genetic properties, no single existing methodology has the scope for accurately genotyping all CNV classes. The dynamic range that exists within complex CNPs poses significant challenges for accurate genotyping. In principle, this likely reflects the greater quantitative differences when detecting deletion copy numbers compared to duplications or multi-allelic loci. Distinguishing four from five diploid copy numbers reproducibly, compared to that of one and two, is difficult using standard methodologies [5]. This clearly poses problems for genotyping in association studies.

Such observations have highlighted the substantial ascertainment bias towards the detection of deletion variants from high resolution genome wide studies (77%) [6], and the limitations of certain techniques for assessing these regions [7]. This is further complicated by the fact that CNVs are enriched (10-fold) for segmental duplications (SDs, defined as sequences >95% and >1 kb in length) making characterization of these regions difficult using current methodologies [8]. Whole Genome Sequencing (WGS) in combination with sequence read depth, allows the analysis of many complex regions of the genome that were excluded from conventional genome wide association studies (GWAS) [9]. Seminal work by Sudmant et al. identified nearly 1000 genes within these regions, ranging from 0 to 48 copies at 3 kb resolution. Such a result expands our knowledge of the "assayable" portion of the genome, but indicates that many highly duplicated regions are yet to be analyzed in studies of disease. It has also been demonstrated that a significant proportion of CNPs residing in SD's are not in Linkage Disequilibrium (LD) with nearby single nucleotide polymorphisms (SNPs) [8]. A study of 192 CNPs, showed that only 40% of those located in SDs had high correlation to nearby SNPs, in comparison to 70% of 892 CNPs in unique regions of the genome [8]. These findings illustrate that for a large number of CNPs, genotypes cannot be imputed through the use of tagSNPs and must be measured directly.

There is continued interest in assessing the relevance of multi-allelic CNPs in complex disease, however, in some cases technical difficulties have impeded the reproducibility of these associations [7,10]. In addition, techniques amenable to genotyping in large scale studies can suffer from poor resolution particularly assignment of integer copy number, meaning that a compromise must be reached between accuracy and cost (Table 1). In general the most accurate techniques are the most labor intensive. Given that large numbers of individuals are required for robust associations and batch effects have the potential to create bias in genotyping, it is necessary to understand the strengths and limitations of methodologies used to analyze multi-allelic CNPs.

1.1. Hybridization-based techniques

1.1.1. Fiber FISH

Fluorescence *in-situ* hybridization (FISH) is a visual technique, typically used to identify chromosomal abnormalities from metaphase or interphase spreads using fluorescent probes. The strength of FISH lies in the direct visualization of DNA copy number at the single cell level. Multi-allelic CNPs, however, can be more difficult to analyze, especially when attempting to resolve tandem duplications. A modified approach, known as Fiber FISH, possesses sufficient resolving power to analyze complex structural rearrangements. The principle of this technique involves the release and fixation of DNA molecules from interphase nuclei onto a slide, with the DNA stretched in a linear fashion through mechanical or gravitational force [11]. The DNA fibers can be hybridized with fluorochrome-labeled DNA probes, producing a characteristic "beads on a string" pattern which is easily distinguishable from background probe signals [12]. Visualization of multiple DNA targets can be achieved using multi-colored probes, which appear as barcode-like signal patterns in the presence of tandem duplications [12]. Simple and complex genomic rearrangements, as well as repetitive sequences, can be accurately resolved using Fiber FISH. A striking example of the ability to resolve complex multi-copy gene re-arrangements was demonstrated in a study of the salivary amylase (*AMY1*) gene [13]. Individuals with tandem duplications of >10 copies could be accurately resolved using Fiber FISH. Techniques which measure changes in diploid dosage have the potential to miss assign complex copy numbers as *de-novo* events in the offspring, purely due to the number of different combinations and inheritance patterns [2]. A key advantage to using Fiber FISH, is that it allows the determination of CN per allele which is important for studies of inheritance and disease [13]. Limitations of

Table 1
Methods to measure complex copy number polymorphisms.

	Fiber FISH	Southern Blot	PFGE	QPCR[b]	MAPH[b]	MLPA[b]	PRT[b]	SNP array	Array CGH[a]	NGS
Detection	Absolute copy number	Inferred absolute copy number/change from diploid dosage	Inferred absolute copy number	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Change from diploid dosage	Absolute copy number
Sample	Cells	2–5 μg DNA	2–5 μg DNA	5–10 ng DNA	0.5–1 μg DNA	100–200 ng DNA	10–20 ng DNA	0.5–1 μg DNA	0.5–1 μg DNA	1–2 μg DNA
Loci	Single	Single	Single	Single	>40	>40	Single	>2 million	>2 million	Genome-wide
Throughput	Low	Low	Low	High	High	High	High	High	High	Low/moderate
Minimum resolution	>1 kb[12]	>1 kb[14]	0.5–1 kb[18]	100 bp	100 bp	100 bp	100 bp	5–10 kb[47]	5–10 kb[47]	>1 kb[9]
Cost per sample	Low	Low	Low	Low	Low	Low	Low	Moderate	Moderate	High
Time to result	>24 h	2–3 days	2–3 days	4 h	>24 h	>24 h	4 h	>24 h	>24 h	2–3 days
Labor requirement	High	High	High	Low	Low	Low	Low	Moderate	Moderate	High

[a] High resolution Array CGH can achieve a minimum resolution of >500 bp[6].
[b] Minimum resolution is in general the length of a single probe.