Genomics 102 (2013) 15–26

Contents lists available at SciVerse ScienceDirect

Genomics

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

Genomic evolution and polymorphism: Segmental duplications and haplotypes at 108 regions on 21 chromosomes

Craig A. McLure [a,b,c,d], Peter Hinchliffe [a], Susan Lester [a,e], Joseph F. Williamson [a,b,*], John A. Millman [a,f], Peter J. Keating [a,b], Brent J. Stewart [a,b], Roger L. Dawkins [a,b,g]

[a] C.Y. O'Connor ERADE Village, PO Box 5100, Canning Vale, Western Australia 6155, Australia
[b] School of Veterinary and Biomedical Sciences, Division of Health Sciences, Murdoch University, Murdoch, Western Australia 6150, Australia
[c] Department of Veterinary Sciences, University of Melbourne, Victoria 3000, Australia
[d] Genetic Technologies Limited, Fitzroy, Victoria, 3065, Australia
[e] Rheumatology Department, The Queen Elizabeth Hospital, Woodville, South Australia, Australia
[f] TAFEWA Swan Campus, Hayman Road, Bentley, Western Australia 6102, Australia
[g] Faculty of Medicine and Dentistry, University of Western Australia, Nedlands, Western Australia 6009, Australia

ARTICLE INFO

Article history:
Received 14 June 2012
Accepted 27 February 2013
Available online 13 March 2013

Keywords:
Ancestral haplotypes
Segmental duplication
Complex disease
Gene Copy Number Variation
Evolution
Conserved polymorphism

ABSTRACT

We describe here extensive, previously unknown, genomic polymorphism in 120 regions, covering 19 autosomes and both sex chromosomes. Each contains duplication within multigene clusters. Of these, 108 are extremely polymorphic with multiple haplotypes.

We used the genomic matching technique (GMT), previously used to characterise the major histocompatibility complex (MHC) and regulators of complement activation (RCA).

This genome-wide extension of this technique enables the examination of many underlying *cis*, *trans* and epistatic interactions responsible for phenotypic differences especially in relation to individuality, evolution and disease susceptibility.

The extent of the diversity could not have been predicted and suggests a new model of primate evolution based on conservation of polymorphism rather than de novo mutation.

1. Introduction

Over half of the observed genetic variation in humans is clustered within genomic regions containing segmental duplications. Interestingly, these polymorphic regions account for only approximately 5% of the genome and tend to be clustered within distinct genomic blocks [1–7]. The principal aim of the present study is to develop a screening test which prospects for biologically important differences and especially those which underlie disease susceptibility and primate evolution. The secondary aim is to determine whether genome-wide polymorphism is sufficient to account for the individuality of humans.

The genomic matching technique (GMT) was developed as an approach to finding suitable bone marrow donors and recipients. After exhaustive testing, the procedure has proven efficient and reliable in recognising alternative polymorphic sequences (haplotypes) within family studies. Identity by GMT predicts a successful transplant outcome [8,9]. To our initial surprise, it transpires that haplotyping is achieved by amplifying duplicated sequences flanked by highly conserved priming sites. In a new application of the technique, here referred to as "duplotyping", we ask how much polymorphism exists in regions of known duplication. After designing a primer pair with the potential to amplify linked duplicons, we tested each pair by comparing the amplification products from different subjects. In this way, we were able to estimate the extent of polymorphism within each duplicated region.

Previous testing of genomic regions, such as the MHC and the RCA, has confirmed the utility of this approach. Multiple amplification products reflect duplicons of varied lengths as happens when different insertions and deletions (indels) accumulate in one copy rather than another. These indels have been shown to be characteristic of each haplotype so that length can be used for haplotyping and for duplotyping.

Differences in the amount of product of a given length relate to the number of duplicated sequences of that length. Thus, duplication can be detected even when the duplicons have the same sequence in *cis* and have not yet accumulated indels (homoduplications).

Duplotyping of the human MHC has already demonstrated the importance of duplication in polymorphic blocks and their relevance to complex disease [10]. Clusters of multicopy gene families [11–14] are distributed throughout ~3.5 megabases (Mb) and were found to

Abbreviations: AH, ancestral haplotype; CEPH, Centre d'Etude du Polymorphisme Humain; GCNV, Gene Copy Number Variation; GMT, genomic matching technique; MHC, major histocompatibility complex; PFB, polymorphic frozen block; RCA, regulators of complement activation.

* Corresponding author at: C.Y. O'Connor ERADE Village, PO Box 5100, Canning Vale, Western Australia 6155, Australia. Fax: +61 8 93971559.
E-mail address: jfw@cyo.edu.au (J.F. Williamson).

0888-7543/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2013.02.011

contain extreme levels of polymorphism including genomic duplications, Gene Copy Number Variations (GCNV)s, retroviral and genomic indels and SNPs [15–17]. Specific combinations of these features, including both coding and non-coding polymorphisms, segregated as nuclear haplotypes through multi-generation families [17]. These haplotypes are precise markers of several hundred kilobases (kb) of sequence [10]. Their occurrence in unrelated individuals implies conservation over many generations and led to the designation ancestral haplotypes (AHs). Recombination occurs between rather than within blocks [10,17–20]. The high polymorphic content and the apparent "freezing" of diverse sequences resulted in these regions being termed "polymorphic frozen blocks" (PFBs) [10,19].

Following the definition of MHC polymorphism, it became possible to investigate the role of genetic susceptibility to diseases. It transpired that particular AHs are associated with specific diseases [10,21,22]. The mechanisms responsible are multifactorial and dependent upon haplospecific interactions of coding and non-coding sequences [23,24]. AHs provide a means of defining these interactions [10,17,20,25] including epistasis.

The extent and importance of ancestral or extended haplotypes was first demonstrated with the identification of haplospecific copy number variations of the complement gene, C4 within the central MHC [26,27]. Each AH has a specific copy number, which relates, in turn, to serum concentration and susceptibility to disease [28].

Although first discovered in the MHC, quantal structure of the genome is now recognised as characteristic of the entire genome [1,29,30], as is the importance of segmental duplications and GCNV on phenotype [31]. Recently high throughput assays such as SNP and Multiplex Ligation-dependent Probe Amplification (MLPA) [32] have been used to detect differences in copy number but with limited success. The MHC and HapMap experiences show that SNP haplotypes of complex regions are misleading. Genomic duplication and especially GCNVs complicate the assignment of SNPs and the determination of phase remains ambiguous until the haplotypes have been assigned independently by demonstrating inheritance by family segregation [20,25,33].

Accordingly, we have developed the genome wide "duplotyping" approach in order to discover new haplotypes directly. The approach relies upon the amplification of multiple polymorphic elements located within linked duplicons, avoiding the risk of inferring haplotypes from independent SNPs and microsatellites. Each of the duplicons has evolved independently from an ancestral sequence. It follows that the specific combinations of duplicons define informative haplotypes efficiently.

Each test requires a single PCR, making the "duplotyping" approach an excellent cost-effective and informative alternative to direct sequencing of multiple individuals. The utility of the technique has been demonstrated over decades of clinical practice [34–38]. Matching GMT profiles of donors and recipients predicts a successful bone marrow transplant [8,9].

GMT has also defined haplotypes in the canine MHC [39], the human RCA [40] and the zebrafish orthologue of human Mannose binding lectin (*MBL2*) [41]. Recently, Lester and colleagues demonstrated an epistatic interaction between the RCA alpha block haplotypes and the MHC in Primary Sjögren's Syndrome [24].

Here we extend the approach to 80 genomic blocks and reveal previously unknown haplospecific polymorphism in humans and in syntenic clusters of other species.

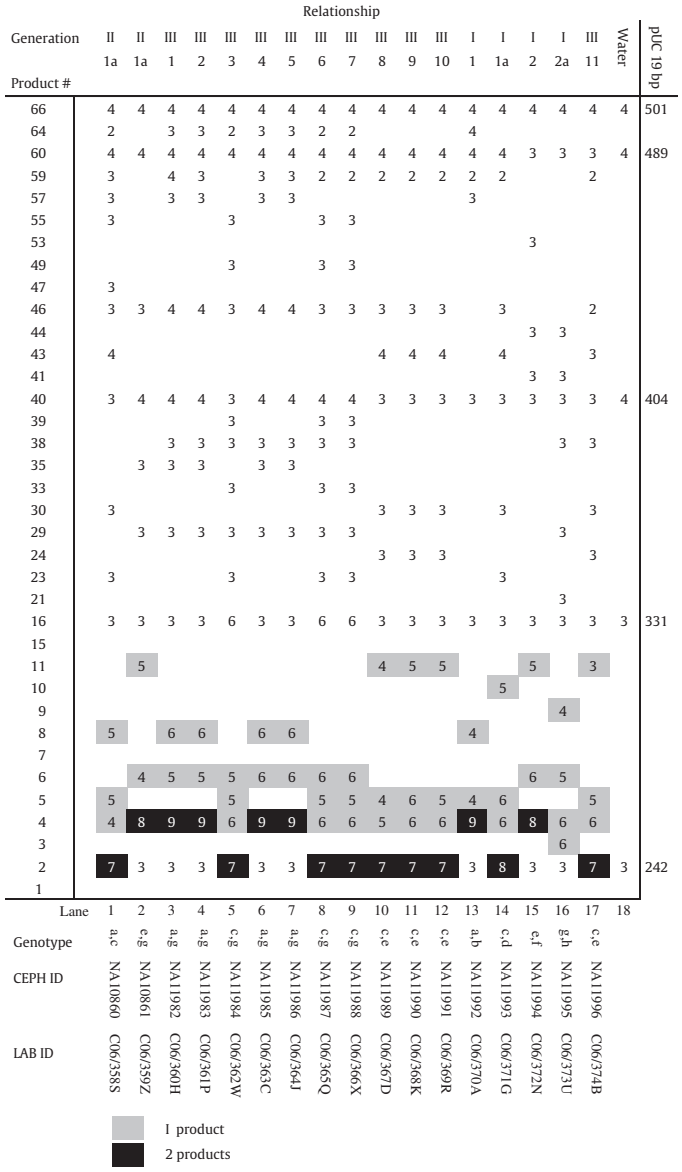
2. Results

2.1. Quantitation and characterization of amplification products

2.1.1. Haplotype analysis in families

An example of the amplification profile is shown in Table 1. In this case, primer pair CYO_5_2 was used to amplify samples from 17 members of a well studied 3 generation CEPH (Centre d'Etude du Polymorphisme Humain) family used to assign individual haplotypes and the resulting composite genotypes throughout the genome.

Table 1
Tabulation and analysis of products from a 3 generation family CEPH Pedigree 1362.



The raw results (shown in Supplementary Fig. 1) are tabulated using an internationally verified and reproducible scoring system, which has been proven to reflect copy number [39,41]. This system allows detection of qualitative and quantitative differences in the amplification products and therefore a precise estimate of polymorphism.

The direct contribution of each haplotype is revealed by comparing the members between and within generations and by demonstrating unequivocal segregation of inheritance.

As shown in Table 1, the grandparents (I1, I1a, I2, I2a) are designated *ab, cd, ef* and *gh* respectively. Their children are *ac* for the father (II1) and eg for the mother (II1a). By inspection of the patterns, it is possible to determine which products are attributable to each haplotype. To confirm these assignments, the patterns in the third generation are examined and the haplotypes are assigned.

The results shown in Table 1 are unequivocal because the family has three of the four possible genotypes (*ag, ae, ce, cg)* in the third generation and each has a different pattern as summarised in Fig. 1.