Gene xxx (2008) xxx-xxx

Contents lists available at ScienceDirect

Gene

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

Analysis of the multi-copy gene family *FAM90A* as a copy number variant in different ethnic backgrounds

Nina Bosch [a,b], Geòrgia Escaramís [a,b], Josep M. Mercader [a,b], Lluís Armengol [a,b], Xavier Estivill [a,b,c,*]

[a] *Genetic Causes of Disease Group, Genes and Disease Programme Center for Genomic Regulation (CRG-UPF), Barcelona, Catalonia, Spain*
[b] *CIBER en Epidemiología y Salud Pública (CIBERESP), Barcelona, Catalonia, Spain*
[c] *Department of Health and Experimental Life Sciences, Pompeu Fabra University (UPF), Barcelona, Catalonia, Spain*

ARTICLE INFO

Article history:
Received 18 March 2008
Received in revised form 30 April 2008
Accepted 2 May 2008
Available online xxxx

Received by M. D'Urso

Keywords:
FAM90A (family with sequence similarity 90, GenBank: NM_018088)
Inter-population variability
Real-time PCR

ABSTRACT

Copy number variants contribute extensively to inter-individual genomic differences, but little is known about their inter-population variability and diversity. In a previous study (Bosch et al., 2007; 16:2572–2582), we reported that the primate-specific gene family *FAM90A*, which accounts for as many as 25 members in the human reference assembly, has expanded the number of *FAM90A* clusters across the hominoid lineage. Here we examined the copy number variability of *FAM90A* genes in 260 HapMap samples of European, African, and Asian ancestry, and showed significant inter-population differences ($p < 0.0001$). Based on the recent study of Stranger et al. (2007; 315:848–853), we also explored the correlation between copy number variability and expression levels of the *FAM90A* gene family. Despite the high genomic variability, we found a low correlation between *FAM90A* copy number and expression levels, which could be due to the action of independent *trans*-acting factors. Our results show that *FAM90A* is highly variable in copy number between individuals and between populations. However, this variability has little impact on gene expression levels, thus highlighting the importance of genomic variability for genes located in regions containing segmental duplications.

1. Introduction

Nucleotide and gene copy diversity is an important driving force of organism evolution and also of selection within any given species (Nguyen et al., 2006; Norton et al., 2007). Genetic variation occurs at multiple levels, from single nucleotide polymorphisms (SNPs), to structural changes such as inversions, deletions or duplications of large DNA fragments, commonly known as copy number variants (CNVs). Recently, it has been shown that BAC array studies have led to an over-estimation of the size of CNVs, and therefore fine-scale analysis of the architecture of these CNVs is necessary (Perry et al., 2008). Although many efforts are needed to elucidate the impact of CNVs on gene expression, recent work has indicated a contribution of both SNPs and CNVs to the variability in the expression of human genes (Stranger et al., 2007).

Genomic copy number gains and losses could have a major impact on gene expression by directly modifying mRNA levels, interrupting a gene transcript, or by having positional effects on regulatory regions (Beckmann et al., 2007; Estivill and Armengol, 2007). Several copy number changes have been found to be related to disease phenotypes, such as the low copy number of *FCGR3B* associated with a predisposition to develop glomerulonephritis (Aitman, 2006) and autoimmune disorders (Fanciulli et al., 2007), whilst a high copy number of *CCL3L1* could have a protective role against HIV infection (Gonzalez et al., 2005) (reviewed in Estivill and Armengol, 2007). Some of these CNVs have been analyzed in different populations and inter-population analyses have revealed a wide variability in copy number (Gonzalez et al., 2005; Perry et al., 2007; Jakobsson et al., 2008). Thus, *CCL3L1* and *AMY1* show striking differences depending on the population examined, indicating that variability in copy number has probably contributed to adaptation to different environments (Gonzalez et al., 2005; Perry et al., 2007). Therefore, it is important to characterize the variability of CNVs amongst different populations and to evaluate their consequences at the expression level.

The generation of multi-copy gene families is an important factor in the generation of biological novelty (Lynch and Conery, 2000). Multi-copy gene families could be generated as a result of unequal crossover processes between segmental duplications (SDs). Several multi-copy gene families, such as olfactory receptor genes or defensin clusters, are located along the distal SDs on chromosome 8p23.1. Copy

Abbreviations: SNP, single nucleotide polymorphism; CNV, copy number variant; SD, segmental duplication; REPD, distal segmental duplications on 8p23.1; REPP, proximal segmental duplications on 8p23.1; UPL, universal probe library; Cp, crossing point.

* Corresponding author. Genes and Disease Programme, Centre for Genomic Regulation (CRG), Plaça Charles Darwin s/n, (Carrer Dr Aiguader, 88), Room 521, PRBB Building, 08003 Barcelona, Catalonia, Spain. Tel.: +3493 316 0159; fax: +3493 316 0099.
E-mail address: xavier.estivill@crg.es (X. Estivill).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.05.003

number variability in the case of defensins has been well established. For alpha defensins *DEFA1/3* variation can range from 4 to 14 tandem array repeats, with 10% of the individuals completely lacking *DEFA3* (Aldred et al., 2005; Linzmeier and Ganz, 2005; Ballana et al., 2007). The beta defensin gene cluster ranges from 2 to 12 copies per diploid genome (Hollox et al., 2003). Recent studies have revealed that individuals with three or less copies of the *DEFB4* gene have a higher risk to develop Crohn's disease (Fellermann et al., 2006). Similarly, individuals with higher genomic copy number of beta defensin genes show association with psoriasis predisposition (Hollox et al., 2008).

Here we analyze the copy number variability of *FAM90A*, a primate-specific gene family, which accounts for as many as 25 members per haploid genome. This family has undergone expansion during primate evolution, and it is located in an extremely dynamic region of the human genome regarding structural variation. We have studied inter and intra-population variability in 260 subjects from three different ethnic backgrounds and have explored the correlation between *FAM90A* copy number and its expression levels.

2. Materials and methods

2.1. Samples analyzed

Samples included in the study were 87 residents from Utah with ancestry from northern and western Europe (CEU); 87 individuals, Han Chinese from Beijing (CHB) and Japanese from Tokyo (JPT); and 86 Yoruba individuals from Ibadan (YRI), Nigeria.

2.2. Sequence similarity searches

To establish the nucleotide identity level between the 25 members of the *FAM0A* gene family, we performed BLAT and BLAST searches using the *FAM90A1* member on chromosome 12p13 as a query. The same method was used to map the four different clusters of *FAM90A* genes located on the 8p23.1 distal duplications (REPD), and the single copies on the proximal duplicons (REPP).

2.3. Real-time PCR analysis

For the quantitative real-time PCR amplification, three sets of universal probe library (UPL) probes and primer pairs were used, one probe targeting the second exon of *FAM90A* genes (UPL probe #79), another probe targeting exon 6 (UPL probe #29) and another targeting the RNAseP, a single copy gene (UPL probe #14) used to normalize *FAM90A* values. Probes and primer sets were designed at the ProbeFinder Design assay center website (https://www.roche-applied-science.com/sis/rtpcr/upl/adc.jsp). Real-time PCR was performed in the LightCycler® 480 System (Roche Molecular Diagnostics), using the following program conditions for both amplicons: 10 min of pre-incubation at 95 °C followed by 45 cycles of 15 s at 95 °C, 1 min at 59 °C and 30 s at 72 °C. For each 10 μl sample in a 384-multiwell plate (Roche Molecular Diagnostics), individual reactions were carried out in triplicate according to the manufacturer's instructions. Independent genomic DNA-based standard curves were used to determine the efficiencies of *FAM90A* target amplifications. Estimates of *FAM90A* copies quantification were obtained in the form of crossing point (Cp) values based on the 'second derivative maximum' method as computed by the LightCycler® 480 (Roche Molecular Diagnostics). Further data analysis was performed with the Cp raw data as described by Pfaffl (2001).

2.4. Statistical analysis

Normalized gene expression values from the three different populations (CEU, CHB/JPT and YRI) were downloaded from the Sanger Genevar webpage (http://www.sanger.ac.uk/humgen/genevar/). Data retrieved by the Illumina™ probe GI_42658889-S, which hybridizes to *FAM90A* members, was used to explore the correlation between *FAM90A* expression levels and *FAM90A* copy number.

Individuals included in the analysis were the same as for the *FAM90A* copy number study. Since the $\log_2$ values corresponding to the two real-time quantitative assays were highly equivalent, we used the average values of the two probes to explore association between gene dosage and expression levels. To determine inter-population differences in relation to *FAM90A* copy number, a linear model was used, including the gender as a confounding variable and a correlation structure among individuals from the same *trio* in an attempt to capture dependence, if any, between parents and their offspring. The correlation model structure is included in the model by specifying the appropriate total variance–covariance pattern matrix. No statistically significant differences between males and females were found. The same approach was used to detect association between expression levels and *FAM90A* copy number. Here we included gender, origin and the interaction between both variables with *FAM90A* copy numbers to see whether the relationship between expression and *FAM90A* copy numbers differed according to inter-population variables and/or gender differences. A correlation structure was also specified to capture dependence among individuals from the same family. None of the interactions were statistically significant as well as gender and origin effects by themselves; therefore results are shown as a simple linear association between expression levels and *FAM90A* copy numbers. All the analysis was performed using the MIXED procedure from SAS System vs. 9.0.

3. Results and discussion

3.1. Sequence similarity and genomic distribution of FAM90A gene family

We have recently characterized a novel multi-copy gene family, *FAM90A*, widely expressed in the 14 tissues we tested, which included lymphoblastoid cell lines (Bosch et al., 2007). The only known *FAM90A* feature is a 19 amino acid motif that resembles a CCHC zinc-finger domain that could be involved in DNA or RNA binding. According to the human reference assembly, the family includes up to 25 members per haploid genome, which share more than 93% identity at the nucleotide level (Fig. 1A). These 25 members are organized as tandem array clusters along the REPD SDs, and as single copies on REPP SDs on 8p23.1, with the exception of *FAM90A1*, located on chromosome 12p13.31 (Fig. 1B). In the case of the HuRef assembly corresponding to Craig Venter sequence, *FAM90A* gene copies map to chromosome 12 (GenBank: CM00473), chromosome 8 (GenBank: CM000469) as well as to chromosome 4 (GenBank: CM000465) (Levy et al., 2007).

The *FAM90A* gene family is only present in primates and it has experienced an increase in copy number, along the hominoid lineage. In our previous study, we observed variability in the number of *FAM90A* members among the 20 human individuals that we tested qualitatively, by pulsed field gel electrophoresis and Southern blotting.

3.2. Intra- and inter-population FAM90A copy number variability

Technical challenges for accurately typing copy number in the *FAM90A* gene family are multiple. Specific assays to determine the presence or the absence for all the 25 *FAM90A* genes in a certain individual are not feasible due to the high level of sequence identity they share. Moreover, absolute quantification using large-scale methods such as multiple ligation probe amplification (MLPA) is not useful as it is not possible to discriminate the large number of gene copies. Thus, we analyzed intra- and inter-population variability for *FAM90A* CNV using a real-time PCR relative quantification method in a sample of 260 HapMap individuals from three different ethnic populations.

Please cite this article as: Bosch, N., et al., Analysis of the multi-copy gene family *FAM90A* as a copy number variant in different ethnic backgrounds, Gene (2008), doi:10.1016/j.gene.2008.05.003

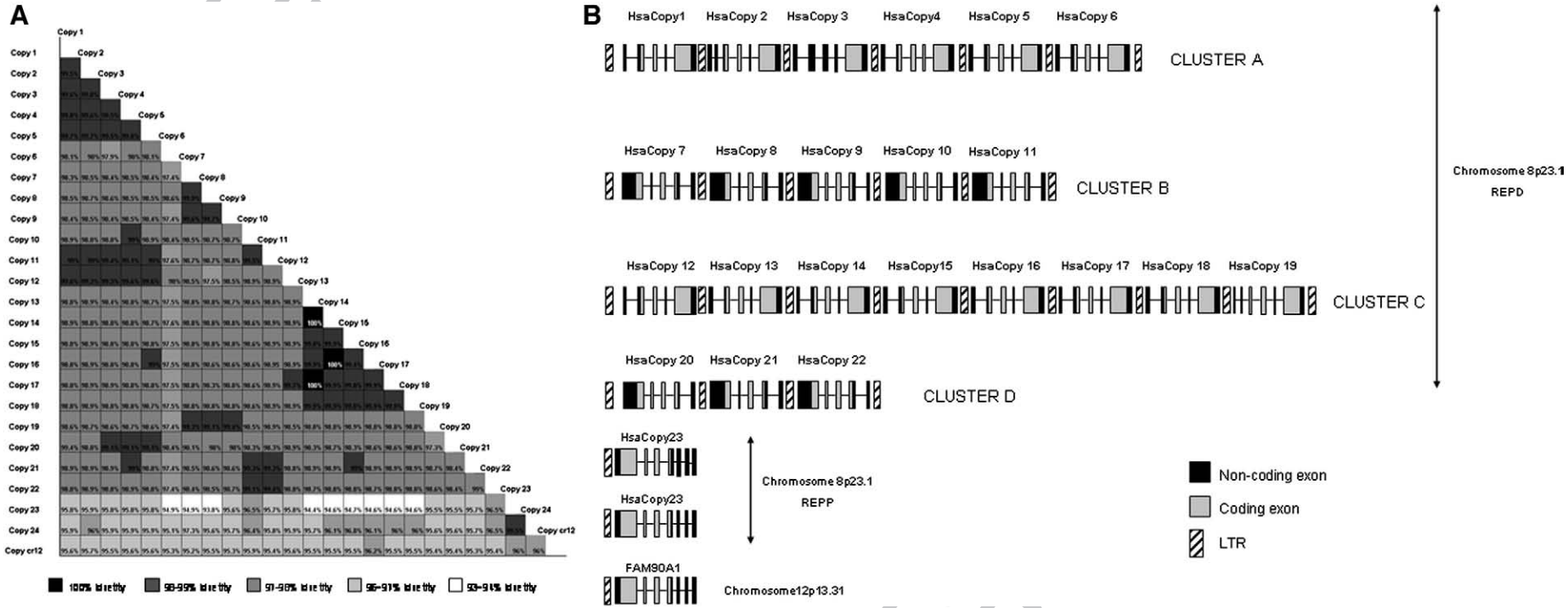

Fig. 1. Nucleotide similarity and distribution of *FAM90A* gene members. (A) Nucleotide identities of the 25 genes (Copy1 to Copy 25) of *FAM90A* family. Taking the *FAM90A1* gene on chromosome 12 as a reference, the minimum identity shared between two different copies is 93%. (B) Based on the current human genome assembly (hg.18) there are 4 different *FAM90A* clusters (A, B, C and D) along the distal duplications on chromosome 8p23.1. In the human reference assembly, these polymorphic clusters contain 6, 5, 8 and 3 *FAM90A* members respectively. The 24 different members on 8p23.1 regions are named HsaCopy1 to HsaCopy24 according to a distal to proximal criteria. Two single copies, HsaCopy23 and HsaCopy24 are found as single copies on the proximal duplicons (REPP). Finally, the single *FAM90A1* member maps to chromosome 12p13.31. Coding and non-coding exons are symbolized as black and grey boxes. The LTR elements that are flanking *FAM90A* members are depicted as striped boxes.

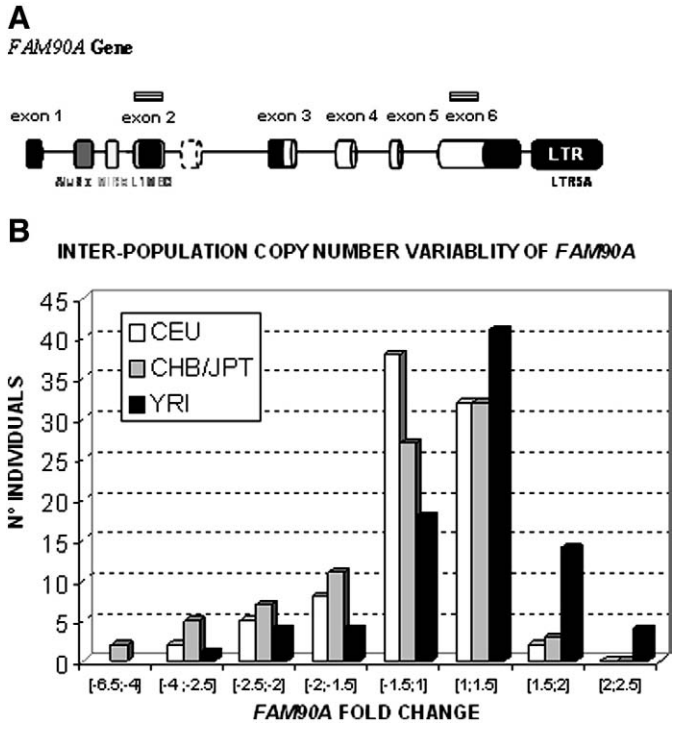

Fig. 2. Analysis of population variability of CNV at the *FAM90A* gene. (A) *FAM90A* gene structure. Filled boxes correspond to non-translated exons and open boxes to the coding sequence. The alternatively spliced exon is represented by a dotted line. Repetitive elements are symbolized as gray and black rectangles. Probes targeting exons 2 and 6 are represented as striped boxes. (B) Variability in *FAM90A* copy number among 260 HapMap individuals. *X* axis represents 8 different intervals according to the fold-change values obtained using the mean of the ΔΔ Cp of UPL probe #29 and UPL probe #79. On the *Y* axis, the number of individuals of the three populations included in each interval is represented.

Two different assays targeting the second and the last exon of *FAM90A* genes were used for quantification analysis (Fig. 2A; Supplementary Table 2). Calculations were based on the NA12892 CEU individual as the average sample regarding *FAM90A* copy number. Inter-population differences are shown in Fig. 2B where the CHB/JPT population shows the lower *FAM90A* copy number and the YRI population shows the higher *FAM90A* copy number and with statistically significant differences in *FAM90A* copy number with respect to CEU and CHB/JPT ($p < 0.0001$).

Real-time PCR using the Universal Probe Library (Roche, Mannheim) revealed that the method is valid and accurate for quantification since the two probes used, successfully replicated the values regarding *FAM90A* relative copy number in all samples (Fig. 3A). Moreover, only the YRI population showed statistically significant intra-population differences regarding *FAM90A*; these results are concordant with previous reports using HapMap samples, where the YRI individuals were the most heterogeneous in comparison to other populations (Foster and Sharp, 2004). We confirm that *FAM90A* constitutes an example of a CNV with dramatic variability among individuals, as the current genome assembly contains up to 25 members of this gene family, and real-time quantitative PCR in HapMap samples varied between a −6.3 and 2.1 fold increase in *FAM90A* copy number, with respect to the reference sample. The analysis of *FAM90A* as a CNV represents a good example of how genomic regions flanked by SDs, such as the 8p23.1 region, can give rise to highly polymorphic CNVs. It is widely accepted that such genomic regions enhance variability of CNVs (Sebat et al., 2004; Sharp et al., 2005), leading to different rearrangements by non-allelic homologous recombination, thus increasing or decreasing the copy number of the intervening sequence.

3.3. Correlation between FAM90A copy number and expression levels

To analyze the possible correlation between *FAM90A* copy number and *FAM90A* expression levels, we utilized transcriptional data from the genome-wide expression arrays analyzed by Stranger et al. (2007); they produced a dataset of gene expression from EBV-transformed lymphoblastoid cell lines from HapMap individuals. Positive correlation was detected between *FAM90A* copy number and expression levels ($r^2 = 0.05$, $p = 0.002$) (Fig. 3B); however, the correlation is weak and population independent, indicating that other factors might be

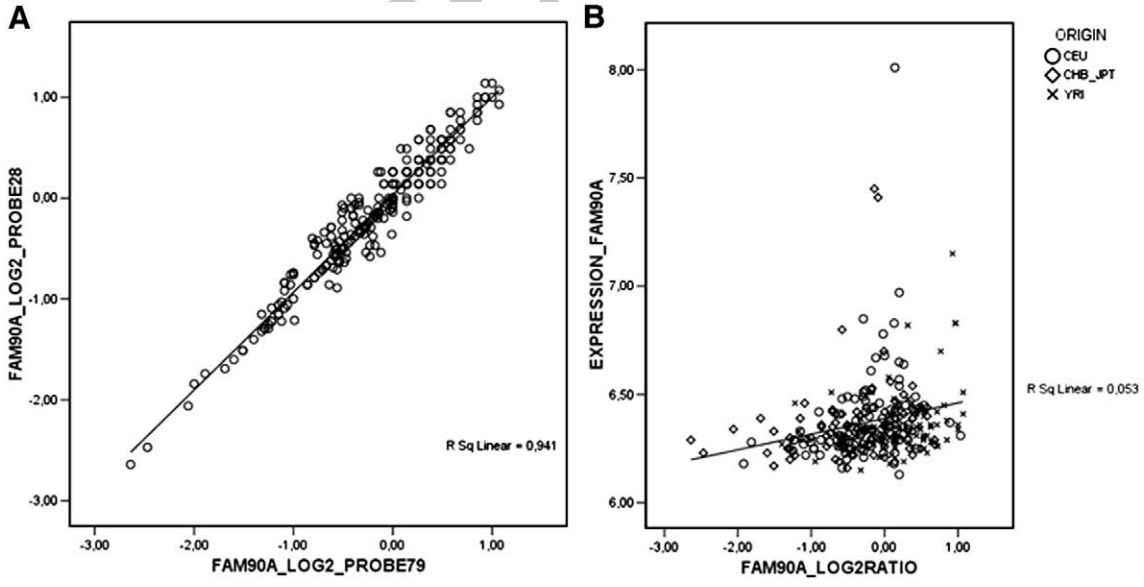

Fig. 3. Correlation between *FAM90A* probes and expression levels. (A) Correlation between the two probes used to quantify *FAM90A* relative copy number. Log2 values for the probe targeting exon 2 are represented on the *X* axis and log2 values for the probe targeting exon 6 are represented on the *Y* axis. (B) Correlation between *FAM90A* copy number variation and *FAM90A* expression levels. In the *X* axis the log2 values corresponding to *FAM90A* copy number are represented. On the *Y* axis there are the *FAM90A* expression levels corresponding to GI_42658889-S probe. The CEU population is represented as circles, the CHB/JPT population as diamonds, and the YRI population as crosses. The regression line between *FAM90A* copy number and expression levels is depicted.

Please cite this article as: Bosch, N., et al., Analysis of the multi-copy gene family *FAM90A* as a copy number variant in different ethnic backgrounds, Gene (2008), doi:10.1016/j.gene.2008.05.003

controlling *FAM90A* expression levels. These results are in agreement with data for other CNVs in this chromosome region. It has also been shown that *DEFA1/3* genes do not have correlation between copy number and expression levels (Aldred et al., 2005). Moreover, whole genome studies have revealed that between 8% and 18% of the heritable variation in transcript levels is due to gene copy number (Stranger et al., 2007), indicating that approximately 80% of this variation is copy number independent. However, this study was based on a limited number of CNVs and also the definition of their boundaries was not precise (Redon et al., 2006). Previous studies have shown that increased transcription levels associated to a reduced copy number are also possible (Stranger et al., 2007). Thus, the low correlation (r^2=0.05, p=0.002) found between gene dosage and expression levels could be due to a complex regulation at the transcription level of *FAM90A* mRNAs. Indeed, heritable *trans*-acting factor variation has been found to be common and is expected to account for a high proportion of gene expression variability between individuals (Morley et al., 2004). Another feasible explanation for the low correlation found, could be that although the HapMap repository represents an enormous resource for genetic studies, lymphoblastoid cell lines might not be the main cell type where expression levels of *FAM90A* are being affected by gene dosage. Similar correlation studies in gene families, which have rapidly expanded in the hominoid lineage, employing appropriate tissues should help to further understand how gene copy number affects transcriptional levels, and therefore its evolutionary constraints.

Acknowledgements

Financial support was received from the Department of Universities, Research and Information Society (2005SGR00008) ("Generalitat de Catalunya"), from the European Union AnEUploidy project (grant number 037627). N. B. is a recipient of a BEFI fellowship from "Instituto de Salud Carlos III FIS-ISCIII"; J.M.M. was supported by the CRG under project SAF2002-00799 (Spanish Ministry of Science and Education) and by fellowship of the Danone Institute (Spain).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2008.05.003.

References

Aitman, T.J., 2006. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439, 851–855.

Aldred, P.M., Hollox, E.J., Armour, J.A., 2005. Copy number polymorphism and expression level variation of the human {alpha}-defensin genes DEFA1 and DEFA3. Hum. Mol. Genet. 14, 2045–2052.

Ballana, E., Gonzalez, J.R., Bosch, N., Estivill, X., 2007. Inter-population variability of DEFA3 gene absence: correlation with haplotype structure and population variability. BMC Genomics 8, 14.

Beckmann, J.S., Estivill, X., Antonarakis, S.E., 2007. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat. Rev. Genet. 8, 639–646.

Bosch, N., et al., 2007. Characterization and evolution of the novel gene family FAM90A in primates originated by multiple duplication and rearrangement events. Hum. Mol. Genet. 16, 2572–2582.

Estivill, X., Armengol, L., 2007. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. PLoS Genet. 3, 1787–1799.

Fanciulli, M., et al., 2007. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat. Genet. 39, 721–723.

Fellermann, K., et al., 2006. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am. J. Hum. Genet. 79, 439–448.

Foster, M.W., Sharp, R.R., 2004. Beyond race: towards a whole-genome perspective on human populations and genetic variation. Nat. Rev. Genet. 5, 790–796.

Gonzalez, E., et al., 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307, 1434–1440.

Hollox, E.J., Armour, J.A., Barber, J.C., 2003. Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. Am. J. Hum. Genet. 73, 591–600.

Hollox, E.J., et al., 2008. Psoriasis is associated with increased beta-defensin genomic copy number. Nat. Genet. 40, 23–25.

Jakobsson, M., et al., 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451, 998–1003.

Levy, S., et al., 2007. The diploid genome sequence of an individual human. Plos Biol. 5 (10), e254.

Linzmeier, R.M., Ganz, T., 2005. Human defensin gene copy number polymorphisms: comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22-p23. Genomics 86, 423–430.

Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequences of duplicate genes. Science 290, 1151–1155.

Morley, M., et al., 2004. Genetic analysis of genome-wide variation in human gene expression. Nature 430, 743–747.

Nguyen, D.Q., Webber, C., Ponting, C.P., 2006. Bias of selection on human copy-number variants. PLoS Genet. 2, e20.

Norton, H.L., et al., 2007. Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. Mol. Biol. Evol. 24, 710–722.

Perry, G.H., et al., 2007. Diet and the evolution of human amylase gene copy number variation. Nat. Genet. 39, 1256–1260.

Perry, G.H., et al., 2008. The fine-scale and complex architecture of human copy-number variation. Am. J. Hum. Genet. 82, 685–695.

Pfaffl, M.W., 2001. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res. 29, e45.

Redon, R., et al., 2006. Global variation in copy number in the human genome. Nature 444, 444–454.

Sebat, J., et al., 2004. Large-scale copy number polymorphism in the human genome. Science 305, 525–528.

Sharp, A.J., et al., 2005. Segmental duplications and copy-number variation in the human genome. Am. J. Hum. Genet. 77, 78–88.

Stranger, B.E., et al., 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315, 848–853.

Please cite this article as: Bosch, N., et al., Analysis of the multi-copy gene family *FAM90A* as a copy number variant in different ethnic backgrounds, Gene (2008), doi:10.1016/j.gene.2008.05.003