Short Communication

Phylogenomic analysis reveals ancient segmental duplications in the human genome

Madiha Hafeez1, Madiha Shabbir1, Fouzia Altaf1, Amir Ali Abbasi*

National Center for Bioinformatics, Program of Comparative and Evolutionary Genomics, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan

A R T I C L E  I N F O

Article history:
Received 21 April 2015
Revised 4 August 2015
Accepted 21 August 2015
Available online 2 September 2015

Keywords:
FGFR
Phylogeny
Gene duplications
Whole genome duplications
Vertebrates
Human chromosomes
Paralogon

A B S T R A C T

Evolution of organismal complexity and origin of novelties during vertebrate history has been widely explored in context of both regulation of gene expression and gene duplication events. Ohno (1970) for the first time put forward the idea of two rounds whole genome duplication events as the most plausible explanation for evolutionizing the vertebrate lineage (2R hypothesis). To test the validity of 2R hypothesis, a robust phylogenomic analysis of multigene families with triplicated or quadruplicated representation on human FGFR bearing chromosomes (4/5/8/10) was performed. Topology comparison approach categorized members of 80 families into five distinct co-duplicated groups. Genes belonging to one co-duplicated group are duplicated concurrently, whereas genes of two different co-duplicated groups do not share their duplication history and have not duplicated in congruency. Our findings contradict the 2R model and are indicative of small-scale duplications and rearrangements that cover the entire span of animal's history.

1. Introduction

One of the major focuses of evolutionary developmental biologists is to unveil the genetic underpinnings of major changes in organismal design and the origin of novelties during evolutionary history of animals. A large body of evolutionary developmental (evo-devo) studies has revealed that differences among closely and distantly related animal taxa are associated with differences in the spatial and temporal aspects of gene expression regulation during development (Villar et al., 2014). Still others consider increase in gene number by duplication as a principle mechanism responsible for increase in organismal complexity and diversity (Ohno, 1973; Van de Peer et al., 2009). Evolution of organismal complexity and origin of novelties during vertebrate history has been widely explored in context of both small and large scale gene duplication events (Abbasi, 2008; Rogers and Gibbs, 2014). The first instance was in the year 1970 when Susumu Ohno put forward the idea of whole genome duplications (WGD) as the most plausible explanation of evolution of form during early history of vertebrates (Ohno, 1970). This notion popularly theorized as “2R hypothesis” (two rounds of WGDs) has been widely debated (Abbasi, 2010a).

Among substantial evidence adduced in favor of ancient vertebrate polyploidy (genome duplications), the most widely cited suggests the existence of paralogons or paralogous genomic segments in vertebrate genomes: homologous chromosomal segments within the genome sharing similar sets of genes (Hwang et al., 2013; Putnam et al., 2008). Precisely, the occurrence of four potential quadruplicated regions, notably on Homo sapiens autosome (Hsa) 1/6/9/19, Hsa 4/5/8/10, Hsa 1/2/8/10 and the HOX-cluster bearing chromosomes Hsa 2/7/12/17, are considered to be structured by two rounds of polyploidy (Furlong and Holland, 2002). However, it is alternatively hypothesized that the profusion of paralogy regions on human chromosomes is due to a high rate of local duplications, translocations and genomic rearrangement events that occurred at widely different time points during early vertebrate history, disputing the central tenet of Ohno’s idea (Hughes et al., 2001).

Over the years, several human paralogons are being surveyed to determine the mechanistic of embarkation of these vertebrate specific paralogy regions. In this regard, HOX and FGFR clusters are most intensively studied paralogons and are considered to have coevolved through 2R-WGD, collectively contributing toward higher complexity in vertebrates (Coulier et al., 1997). The earlier studies on these paralogons incorporated data from sparse sampling of genes and were limited to few species, leaving this longstanding mystery yet to be solved (Abbasi, 2008). To dig deeper into the debate, our group is continuously putting the
efforts in assembling and dating the primordial duplication events that molded animal lineage (Abbasi, 2010b; Abbasi and Hanif, 2012; Ajmal et al., 2014; Ambreen et al., 2014; Asrar et al., 2013). Previously we analyzed the evolutionary histories of 21 multi gene families (93 genes) with triplicated or quadruplicated distribution on human FGFR bearing chromosomes (4/5/8/10). To perform the robust phylogenetic analysis, 1494 sequences were obtained from diverge range of vertebrate and invertebrate species. These data was not substantiated in the favor of en bloc duplications, instead, it appeared that paralogy blocks residing on human FGFR bearing paralogon are the consequence of small-scale duplication events that spread across the entire history of animals (Abbasi, 2010b; Abbasi and Grzeschik, 2007; Ajmal et al., 2014; Ambreen et al., 2014; Asrar et al., 2013; Martin, 2001).

In the present study, we extended our previous work (Ajmal et al., 2014) and investigated the evolutionary history of further 59 human multigene families with three or fourfold representation on FGFR bearing chromosomes (Hsa 4/5/8/10). A thorough phylogenetic analysis of these gene families was performed by employing the currently available well-annotated and high-quality finished genomic sequence data using neighbor joining (NJ) and maximum likelihood (ML) methods. In addition, congruency among topologies of 80 phylogenies (59 present data and 21 previous data) was scrutinized to identify the genes which could have duplicated simultaneously at the root of vertebrate history (Abbasi, 2010b; Ajmal et al., 2014) (Fig. 1).

2. Materials and methods

2.1. Dataset

Genes from 59 gene families (in total 447 genes and 4807 amino acid sequences) with threefold or fourfold representations on human FGFR bearing chromosomes were included in the analysis. The chromosomal location of human gene families was obtained from Ensembl genome browser (Hubbard et al., 2002;
Paul et al., 2014), 11 of these families have members on each of the human FGFR bearing chromosomes while 48 have their members on at least three of those chromosomes (Fig. 1a and Appendix 1). Information regarding molecular functions (Appendix 1) of selected gene families was retrieved from GeneReport, available at SOURCE (Diehn et al., 2003).

BLASTP (Altschul et al., 1990), using bidirectional best hit strategy, was employed to attain closest putative orthologous sequences of the human proteins in other species (Clamp et al., 2003; Hubbard et al., 2002). For those organisms for which sequences was not available at Ensembl (Paul et al., 2014), the BLASTP (Altschul et al., 1990) search was carried out against the protein database available at the National Center for Biotechnology Information (Johnson et al., 2008) and the Joint Genome Institute (Nordberg et al., 2014).

The list of sequences used in the analysis (from 41 species including 25 tetrapods, 5 teleost fish and 11 invertebrates) is provided in Supplementary material (Appendix 2). The species that were selected in the analysis comprises of Homo sapiens (Human), Mus musculus (Mouse), Pan troglodytes (Chimpanzee), Gorilla gorilla (Gorilla), Callithrix jacchus (Marmoset), Pongo abelii (Orangutan), Macaca mulatta (Macaque), Rattus norvegicus (Rat), Oryctolagus cuniculus (Rabbit), Gallus gallus (Chicken), Taeniopygia guttata

Fig. 1 (continued)
Fig. 2. The relative timing of duplication events that expanded the multigene families residing on human FGFR paralgon. For 80 multigene families (in total 536 human genes) residing on Hsa 4/5/8/10, in total 201 duplication events were detected before vertebrate–invertebrate split and 263 duplications were detected after vertebrate–invertebrate and before tetrapod–bony fish divergence. Only 15 tetrapod specific duplication events were detected.
(Zebra finch), Canis familiaris (Dog), Felis catus (Cat), Bos Taurus (Cow), Equus caballus (Horse), Loxodonta africana (Elephant), Dasyurus novemcinctus (Armadillo), Myotis lucifugus (Microbat), Pteropus vampryus (Megabat), Monodelphis domestica (Opossum), Ornithorhynchus anatinus (Platypus), Anolis carolinensis (Lizard), Pelodiscus sinensis (Chinese softshell turtle), Xenopus tropicalis (Frog), Eri-
naceus europaeus (Hedgehog), Danio rerio (Zebrafish), Takifugu rubripes (Fugu), Tetraodon nigroviridis (Tetraodon), Gasterosteus aculeatus (Stickleback), Oryzias latipes (Medaka), Ciona intestinalis (Ascidian), Ciona savignyi (Ascidian), Branchiostoma floridae (Amphioxus), Strongylocentrotus purpuratus (Sea urchin), Drosophila melanogaster (Fruit fly),Apis mellifera (Honey bee), Anopheles gambiae (Mosquito), Caenorhabditis elegans (Nematode), Nematos-
tella vectensis (Sea anemone), Hydra magnipapillata (Hydra) and Amphimedon queenslandica (Sponge).

2.2. Alignment and phylogenetic analysis

Amino acid sequences were aligned by using CLUSTAL W (Thompson et al., 1994) under default parameters. Phylogenetic trees were constructed by using Neighbor-Joining (NJ) method (Russo et al., 1996; Saitou and Nei, 1987) with p-distance as amino-acid substitution model, implemented in MEGA version 5 (Kumar et al., 2008; Tamura et al., 2011). Complete deletion option was selected to eliminate any site which can introduce a gap in the sequences. The sequences that were too diverged, disrupting the entire tree were excluded. Tree topologies were supported by bootstrap (Felsenstein, 1985) with 1000 replicates. To systematically check and validate trees with different reconstruction method; Maximum Likelihood with Whelan and Goldman (WAG) model (Whelan and Goldman, 2001), using MEGA 5 program (Tamura et al., 2011) was implemented. Results from both the methods are presented here (Appendix 3 and 4). Gene duplication events were estimated by the branching order of each gene family within the phylogenetic tree (Appendix 5). The tree topology of each gene family was compared with those of other families to test for consistencies in duplication events.

Among the topologies of 59 gene families, the phylogenetic trees of five gene families (CCN, CPEB, HTRA, MARCH, NRG) were rooted with both invertebrate and vertebrate sequences. The phylo-
geneties of 27 families (ABLIM, ACSL, ADAM, ADAMTS, ADRA, ANXA, ARAP, CAMK, COL, DNAJ, DLX, DUSP, EGR, GRK, HNRNP, MAPK, NIPAL, NKPX, PDE, PKD, RHO, RUFY, SEC24, SEPT, SLC39A, TSPAN, UNC) consisted of subfamilies, each of which served to root the other. The phylogenetic trees of DDK, DOCK, EBF, EIF4EBP, FGF, GPR, GPRIN, GRI, KCNIP, MYOZ, NDST, NPM, POLI, PP3C, PSD, SLIT, SNC, SPICK, SYNPO, TCF7 and VDAC were rooted with orthologous sequences from invertebrates. Sequences of vertebrates of DND1 served to root RBM phylogeny. PALLD gene family is rooted with vertebrates of MYOT, sequences of vertebrates of KLF17 served to root KLF phylogeny, PPRGC gene family is rooted with vertebrates of PPRGC1A, SLC26A gene family is rooted by vertebrates of SLC26A8 and TACC phylogeny is rooted with vertebrates TACC3.

3. Results and discussion

In order to test Ohno’s prediction of two rounds of whole genome duplications that were speculated to cast a major impact on ancestral vertebrate genome architecture, we conducted a robust phylogenetic analysis of 59 novel gene families (447 human genes and 4807 sequences in total) (Fig. 1a). These data combined with our previous investigation have categorized the phylogenies of 80 families (with three or four fold representation on human FGFR bearing chromosome Hsa 4/5/8/10) (Ajmal et al., 2014) into five discrete co-duplicated groups (Fig. 1b and Appendix 5). Gene families belonging to a particular co-duplicated group share similar evolutionary history and might have originated through gene cluster duplication event at the root of vertebrate lineage, whereas the genes belonging to different co-duplicated groups may not share the evolutionary history and might not have duplicated simultaneously (Fig. 1b).

Interestingly, congruent and symmetrical topologies of the type ((A)(B))(C) were recovered for the PPP2R2, REEP, FGFR and TCF7 (Fig. 1b) (Ajmal et al., 2014), seeming to be ohnologue families (stemmed by WGD events). We assert, however, that sub-chromosomal duplications and rearrangements might be a more realistic explanation for such congruent and symmetrical tree topology patterns (Abbasi and Hanif, 2012). Furthermore, identifying the timing of duplication events relative to radiation of major animal taxa deciphered approximately similar proportion of duplications before and after vertebrate-invertebrate split (Fig. 2). Thus, relative dating discounts the contention that a burst of gene duplication activity took place in the early vertebrate history after the divergence of invertebrate lineage (McLysaght et al., 2002). These results are in accordance with the data confirming recent segmental duplications (rSDs) in primates (Rogers and Gibbs, 2014) bolstering the hypothesis of a continuous wave of small-scale doublings during animals’ history (Hughes et al., 2001). Therefore, it appears that current hierarchy of human pro-
teome is created by small-scale events, scattered at different times over the evolutionary history of animal’s life.

4. Conclusion

To elucidate the nature of evolutionary events that resulted in the origin of vertebrate quadruplicated or triplicated paralogy blocks, we investigated the evolutionary history of 540 human genes (in total 80 gene families) residing on human FGFR bearing chromosomes (Hsa 4/5/8/10). Our results indicate that ancient segmental duplications (aSDs) and rearrangements helped fuel the key processes involved in shaping the contemporary architecture of human genome. These data compels to infer that the mechanism underlying ancient vertebrate genome evolution is not different from those that occurred during their recent history. Therefore, we are safe to say that tetraploidization is an unlikely event to be proposed as potential shaping force of vertebrate biology and genome architecture.

Authors’ contributions

AAA conceived the project and designed the experiments. M. H., M.S. and F.A. performed the experiments. A.A.A., M.H., M.S. and F.A. analyzed the data. A.A.A., M.H., M.S. and F.A. wrote the paper.

Competing interest statement

The authors declare that they have no competing financial interests.

Acknowledgment

This work was supported by higher education commission of Pakistan.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2015.08.019.


