Linkage disequilibrium analysis of global populations confirms presence of regulatory SNP rs11615992 of human P2RX7 gene and uncovers rs61083578 as potential alternative in fixed allele populations

Mauro Chavez (mchavez8@jh.edu)
Johns Hopkins University https://orcid.org/0000-0001-9126-8608
Jacquelyn Wilson
Johns Hopkins University
Lynnae Racette
Johns Hopkins University
Gregory Crawford
Johns Hopkins University

Short report

Keywords: P2RX7, Allele-specific expression (ASE), linkage disequilibrium (LD), single nucleotide polymorphism (SNP), rs3751143, rs11615992, rs61083578

Posted Date: May 14th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-27709/v1

Abstract

The *P2RX7* gene is ubiquitously expressed throughout the human body and encodes for ligand-gated ion channels. *P2RX7* also contains the non-synonymous single nucleotide polymorphism (SNP) rs3751143 which has been linked to decreased ion channel function and diseases such as prostate cancer, tuberculosis, bipolar disorder, inflammation and Parkinson's disease amongst others. While there have been many studies on this gene, few dive into the allele-specific expression (ASE) of *P2RX7*. Previous research has shown over-expression of the rs3751143 wild-type A allele over the C allele in lung tissues of heterozygous individuals correlated with the presence of downstream SNP variant rs11615992. Linkage disequilibrium was demonstrated between these two SNPs in populations in China. This article aims to expand current knowledge for which SNPs near the *P2RX7* gene are in linkage disequilibrium with rs3751143; expanding upon findings of Peng et al (2019). By using various bioinformatic tools on variant data gleaned from the 1000 Genomes Project, this analysis confirms that the downstream SNP rs11615992 is in strong LD with rs3751143 in populations across the world, outside of the initially studied Chinese groups. Furthermore, in the case of African populations, where rs11615992 seems fixed, rs61083578 was identified as a potential downstream regulatory element to P2RX7 allele-specific expression. The goal of this study is to contribute to the understanding of *P2RX7* gene expression on a global scale with respect to regulation by SNPs surrounding the *P2RX7* gene itself, strengthening the relevance of the research performed by Peng et al (2019).

Introduction

Purinergic receptor P2RX7 (P2RX7, NCBI Gene ID: 5027) is a protein coding gene that produces an inotropic class of ligand-gated ion channels and is expressed in many tissues throughout the human body; though mostly expressed in brain and skin cells (NCBI 2020). The protein encoded from *P2RX7* differs from other P2X receptor genes, as well as other ligand-gated channels, since it can transport molecules up to 900 Da (Benzaquen et al 2019). The P2RX7 protein has been shown to be associated with many biological processes, and is a potential target for anti-inflammatory therapy (Peng et al 2019). As *P2RX7* is expressed in many cell lines, including immune and non-immune cells, and is involved with genetic expression signaling, macrophage immune response, and membrane voltage potential of both the cell plasma membrane and the nuclear membrane (Fuller et al 2009). Numerous splicing variants of this gene have been identified, many of which seem to be involved with nonsense-mediated decay (NCBI 2020). Furthermore, single nucleotide polymorphisms in *P2RX7* have been shown to play a role in a myriad of diseases, ranging from mood disorders, bipolar disorder, and unipolar depression (McQuillin et al 2009) to tuberculosis susceptibility (Peng et al 2019).

Previous research has been performed on many different SNPs within *P2RX7*. One of the most highly researched SNPs is rs3751143, which was also associated with multiple phenotypes including IL-8 level in serum, a chronic obstructive pulmonary disease biomarker, Toxoplasma gondii killing, and mineral density (Peng et al 2019). ASE results from the lung tissue of 48 donors in China found the rs3751143 wild-type A allele to be overexpressed over the mutant C allele in heterozygous individuals. The

expression ratio (C/A) had a mean ± SD of 0.84 ± 0.16 with 95% confidence interval 0.76–0.92; P = 0.001 (Peng et al 2019). This presented the possibility that SNPs in linkage disequilibrium with rs3751143 play a role in the regulation of the expression of *P2RX7*. Peng et al (2019) focused their study on Chinese populations, performing computational LD analysis on Han Chinese in Beijing (CHB) and Southern Han Chinese (CHS) populations.

By analyzing variants in linkage disequilibrium with rs3751143, Peng et al (2019) uncovered SNP rs11615992. This variant became the focus of many subsequent investigations. Peng et al (2019) deployed a long-distance gene interaction model that inserted the SNP rs116155992 into a region upstream from the *luciferase* gene in a pGL3-promoter vector. Their results demonstrated an increase in expression of luciferase when spliced with both SNP alleles, specifically the A allele of rs11615992 and the A allele of rs3751143 (Peng et al, 2019). This resulted in a more than 40% increased luciferase activity when compared with the rs11615992 G allele (Peng et al 2019). To test whether the SNP rs11615992 may interact with the P2RX7 transcription factor (TF), Peng et al (2019) computationally predicted a TF binding to the area surrounding rs11615992 by using the TRANSFAC database. The results indicated that the TF POU2F1 (POU class 2 homeobox 1) also binds to the region surrounding rs11615992 (Peng et al 2019). Subsequently, this related transcription factor, POU2F1, was used in a ChIP-seq experiment which showed the region associated with rs11615992 displayed significantly higher levels of immunoprecipitation of the antibody for POU2F1 than IgG (Peng et al 2019). This validated that, despite being found downstream of the *P2RX7* gene, rs11615992 has the potential to play an important regulatory role in the transcription of *P2RX7* (Peng et al 2019).

This current study aims to computationally investigate the region surrounding P2RX7 for other SNPs in linkage disequilibrium with rs3751143 in more geographically diverse populations. It is important to note that Peng et al's (2019) chromosome conformation capture assay verified that the region surrounding rs11615992, a position 5 kb downstream of rs3751143, could interact with the *P2RX7* promoter as an enhancer. This leaves open the possibility for both upstream and downstream SNPs in LD with rs3751143 to affect *P2RX7* expression as shown with rs11615992.

Methods

Population Selection

Given that the analysis by Peng et al (2019) focused on two populations in China, Han Chinese in Beijing (CHB) and Southern Han Chinese (CHS), the scope of this analysis was broadened to include a population from each of the following regions: South Asia (SAS), Europe (EUR), East Asia (EAS), America (AMR), and Africa (AFR). The CHB and CHS data sets were included in this workflow to test the reproducibility of the analysis performed by Peng et al (2019). Ensembl variation data from phase 3 of the 1000 Genomes Project for two target SNPs (rs3751143 and rs11615992) was surveyed to select populations from each region displaying comparatively larger minor allele frequencies for both variants in an effort to further diversify the number of variant pairs used in linkage disequilibrium analysis. A

summary of populations used in this analysis reporting the major and minor allele frequency of both SNPs of interest (rs3751143 and rs11615992) can be found in Table 1. The Gambian in Western Division (GWD) population of Africa was intentionally selected due to there being no presence of the rs11615992

minor allele. This population would serve as a control group as no LD with rs3751143 and rs11615992 would be possible.

Table 1
Major and minor allele frequency of SNPs of interest in populations used for analysis

Population	Population code	Region	rs3751143 major allele frequency (A)	rs3751143 minor allele frequency (C)	rs11615992 major allele frequency (A)	rs11615992 minor allele frequency (G)
Han Chinese in Beijing, china	CHB	EAS	0.748 (154)	0.252 (52)	0.767 (158)	0.233 (48)
Han Chinese South	CHS	EAS	0.738 (115)	0.262 (55)	0.771 (162)	0.229 (48)
Kinh in Ho Chi Minh City, Vietnam	KHV	EAS	0.702 (139)	0.298 (59)	0.722 (143)	0.278 (55)
Sri Lankan Tamil in the UK	STU	SAS	0.593 (121)	0.407 (83)	0.672 (137)	0.328 (67)
Toscani in Italy	TSI	EUR	0.715 (153)	0.285 (61)	0.752 (161)	0.248 (53)
Puerto Rican in Puerto Rico	PUR	AMR	0.817 (170)	0.183 (38)	0.837 (174)	0.163 (34)
Gambian in Western Division, The Gambia	GWD	AFR	0.903 (204)	0.097 (22)	1.00 (226)	0.00 (0)

Genome analysis

1000 Genomes phase 3 variant data for chromosome 12 across all populations was downloaded from the 1000 Genome Project website (http://www.internationalgenome.org/) as a VCF file (Auton et al 2015). Variant data was filtered down to a 100 kb window surrounding rs3751143 (50 kb upstream and downstream) to a data set of 3,240 SNPs. This window size was selected as it is just under double the size of the *P2RX7* gene (~ 54 kb). From there, sample lists for each population of study were downloaded as TSV files. These two data sets were merged using python code to generate input files for ldSelect (Nickerson 2004) or Genome Variation Server (http://gvs.gs.washington.edu/GVS150/). Additionally, in the generation of input files, data on the pairing of SNP alleles was also summarized for later queries. Table 2 presents an overview of the number of samples provided by the 1000 Genomes Project for each

population of study and the total number of samples found within the chromosome 12 phase 3 variant data set.

Table 2
Samples available via 1000 genomes project and number of samples identified with phase 3 variants

Population	Number of samples reported in 1000 Genomes project	Number of samples found in phase 3 variant data set
CHB	112	103
CHS	171	105
KHV	124	99
STU	128	102
TSI	112	107
PUR	150	104
GWD	280	113

Regulation analysis

In order to computationally determine the potential regulatory effects of SNPs, NONCODE (Zhao 2016) was utilized to search for SNP proximity to non-coding RNA. Additionally, the MATCH web-app (for querying the TRANSFAC database (Matys 2006)) was used to search for nearby transcription factor binding sites.

Results

Linkage Disequilibrium pattern near P2RX7 gene in global populations

1000 Genomes Project data for the region surrounding *P2RX7* in the populations described above were analyzed to identify SNP(s) in LD with rs3751143. For reference, Peng et al (2019) reported rs11615992 as the only SNP with strong LD to rs3751143 in their two populations of study (r^2 = 0.900 in CHB and r^2 = 0.835 in CHS). In filtering for SNPs with strong LD to rs3751143, a minimum r^2 of 0.800 was used as a score threshold. LD analysis utilizing the Genome Variation Server confirmed that rs11615992 is the only SNP with strong LD to rs3751143 in all populations, excluding the control (GWD) and Puerto Rican in Puerto Rico (PUR). In PUR, no SNPs were in significantly strong LD with rs3751143 as the r^2 of rs3751143 being in LD with rs11615992 fell below the significance threshold to a value of 0.713. In the GWD control population, LD between rs3751143 and rs11615992 was impossible due to the lack of individuals surveyed for variants displaying the rs11615992 minor allele. Nevertheless, in analysis of the GWD data set, SNP rs61083578 was seen to be in strong LD with rs3751143 (r^2 = 0.807). Table 3 captures r^2 linkage disequilibrium scores between rs3751143 and rs11615992 as well as between rs3751143 and

rs61083578. Cells containing N.I. (Not Identified) report that the allele pair did not appear in the results of linkage disequilibrium analysis. Two results of note are the scores of the two populations described in the Peng et al (2019) article. This analysis was able to recover the same r^2 values that were previously reported.

Table 3
LD of rs3751143 with two significantly linked downstream SNPs per population (N.I. = Not Identified).

Population	r^2 of rs3751143 LD with rs11615992	r^2 of rs3751143 LD with rs61083578
CHB	0.900	N.I.
CHS	0.835	N.I.
GWD	N.I.	0.807
KHV	0.906	N.I.
STU	0.713	N.I.
PUR	0.874	N.I.
TSI	0.826	N.I.

Allele pair types in strong linkage disequilibrium

In an effort to better understand the nature of rs3751143 LD with rs11615992, the counts of each variant pair were summed to determine which alleles for each SNP appeared in pairs. Peng et al (2019) observed that “the wild-type A allele of rs3751143 was in strong LD with A of rs11615992, while the mutant one C of rs3751143 was in LD with G of rs11615992”. This pattern remained consistent across all 6 populations that displayed both alleles for each SNP as shown in Table 4 which shows the count of per strand genotypes for allelic variant pairs of rs3751143 and rs11615992 across all populations of study.

Table 4
Genotype paring counts per population between rs3751143 and rs11615992

Population	rs3751143 w.t. (A) with rs11615992 w.t. (A)	rs3751143 w.t. (A) with rs11615992 mutant (G)	rs3751143 mutant (C) with rs11615992 w.t. (A)	rs3751143 mutant (C) with rs11615992 mutant (G)
CHB	154	0	4	48
CHS	154	1	8	47
KHV	139	0	4	55
STU	121	0	16	67
TSI	153	0	8	53
PUR	170	0	4	34
GWD	204	0	22	0

The GWD population variant data, being absent of rs11615992 minor allele, did not present LD between that SNP and rs3751143. However, as noted above, this population revealed that rs3751143 was in strong LD with rs61083578. The counts of these two allele variant pairs closely resembles those found in the other data sets as shown in Table 5 which captures the count of per strand genotypes for allelic variant pairs of rs3751143 and rs61083578 across all populations where the two were in strong LD. TSI and PUR populations both had a single case where mutant rs3751143 was paired with mutant rs61083578 but no LD was found. All other populations only showed wild-type pairings.

Table 5
Genotype paring counts for GWD population between rs3751143 and rs61083578

Population	rs3751143 w.t. (A) with rs61083578 w.t. (C)	rs3751143 w.t. (A) with rs61083578 mutant (T)	rs3751143 mutant (C) with rs61083578 w.t. (C)	rs3751143 mutant (C) with rs61083578 mutant (T)
GWD	202	2	2	20

Potential regulatory effects of rs61083578

The region surrounding rs61083578 was surveyed for potential regulatory elements that might contextualize the relevance of Peng et al's (2019) ASE findings (that the A allele of rs3751143 is over-expressed relative to the C) to the GWD population. NONCODE data viewed in the UCSC genome browser placed rs61083578 inside lnc-RNA (ID: NONHSAT232452.1). Furthermore, the TRANSFAC database reported a potential binding site for CCAAT Enhancer Binding Protein alpha in close proximity to rs61083578.

Discussion

In their original article, Peng et al (2019) identify the novel regulatory SNP rs11615992 as a driving force behind the observed allele-specific expression of the clinically significant *P2RX7* gene. The goal of this analysis was to study additional populations, from different geographical regions, in an attempt to uncover the same linkage disequilibrium between rs3751143 and rs11615992; a key characteristic of the *cis*-regulatory capability of rs11615992. In doing so, the two populations of focus in the Peng et al (2019) article were reanalyzed for LD to ultimately uncover the same results. Furthermore, five of six additionally studied populations uncovered the same LD pattern with strong r^2 scores. The Sri Lankan Tamil in the UK (STU) population fell below the 0.800 cut-off (with an r^2 score of 0.713), which may be explained by the relative closeness in frequency of the rs3751143 major and minor allele frequencies (0.593 and 0.407 respectively). The LD pattern of SNP specific alleles for each variant discovered by Peng et al (2019) (the wild-type of rs3751143 in LD with wild-type of rs11615992), remained consistent across all data sets, where the minor allele of rs11615992 was present. Finally, in all but one population, no other SNPs were found to be in LD with rs3751143.

Interestingly, in the GWD population, a new SNP was uncovered to be in strong LD with rs3751143 in the absence of rs11615992. This data set contains a fixed A allele for rs11615992, similar to the YRI African population as noted in Peng et al's paper. In the GWD population, rs3751143 was in strong LD with rs61083578 ($r^2 = 0.807$), a variant ~ 14 kb downstream. This is almost triple the distance between rs3751143 and rs11615992. Nonetheless, there is reason to believe rs61083578 potentially stands in as a proxy for rs11615992 as a cis-regulatory factor of *P2RX7* expression. In the Peng et al (2019) publication, they present 3C assay data that supports rs11615992 interacting with the promoter region of *P2RX7* as shown by high interaction frequency between the P2RX7 promoter region and randomly selected chr12:12163138 position. This region randomly selected for their assay is 4050 bases downstream of rs11615992 and 5409 bases upstream of rs61083578. To quote Peng et al (2019) summarizing their results, "When one-sample t-test was used to roughly compare the ligation efficiency, a significant deviation was obtained ($P < 10^{-6}$), thus indicating that *P2RX7* promoter could interact with the enhancer [chr12:12163138 region in close proximity to rs11615992] and *P2RX7* should be the target gene of this enhancer". This strengthens the possibility that the *P2RX7* promoter region might interact with rs61083578, given that the region used in the 3C assay was nearly as close to rs11615992 as it was to rs61083578. There are additional opportunities for rs61083578 to have regulatory effects on *P2RX7*, considering it is near a CCAAT Enhancer Binding Protein alpha binding site and contained within an annotated long non-coding RNA region.

Conclusion

The goal of this analysis was to expand upon the relevance of the thorough and extensive research conducted by Peng et al (2019) on allele-specific expression of *P2RX7* explained by the downstream SNP rs11615992 being in LD with intronic non-synonymous SNP rs3751143. In their initial paper, analysis was focused on the study of Chinese populations. In an effort to broaden the scope of their findings, this analysis has confirmed the relevance of their discovery in Europe, America, and South Asia, by uncovering

comparably strong LD between these two variants in populations from these regions. Furthermore, in the case of the selected African population, SNP rs61083578 was discovered to be in strong LD with rs3751143. This downstream variant shares some characteristics to rs11615992, in that it sits near a predicted transcription factor binding site, and a region used in Peng et al's (2019) 3C assay which showed potential for interaction between downstream enhancer regions and the *P2RX7* promoter. While the scope of the analysis submitted in this paper is confined to inquiry and predictions performed in a purely computational environment, it opens the possibility for clinically relevant rs3751142 allele specific expression of *P2RX7* to not only be influenced by the downstream SNP rs11615992 but also further downstream SNP rs61083578.

Abbreviations

African- AFR

American-AMR

Allele-specific expression- ASE

East Asian- EAS

Gambian in Western Division- GWD

Han Chinese in Beijing- CHB

Linkage disequilibrium- LD

Puerto Rican in Puerto Rico- PUR

Single nucleotide polymorphism- SNP

Southern Han Chinese- CHS

South Asian- SAS

Sri Lankan Tamil in the UK- STU

Transcription factor- TF

Declarations

Ethics approval and consent to participate

The 1000 Genomes data is made available according to the Fort Lauderdale Agreement.

Consent for publication

The 1000 Genomes data is made available according to the Fort Lauderdale Agreement.

Availability of data and materials

The datasets supporting the conclusions of this article are available in the 1000 Genome repository, http://www.internationalgenome.org/.

Competing interests

The authors declare that there are no competing interests.

Funding

Not Applicable

Authors' contributions

MC collected and managed the data, wrote code for data filtering and reformatting, and submitted it to online resources for analysis. MC also interpreted the results from said analysis. JW and LR contributed background research on gene significance and provided feedback on analysis methodology and results. GW edited and reviewed the manuscript.

Acknowledgements

We would like to acknowledge our professors Sajung Yun Ph.D. and Sijung Yun Ph.D. for providing guidance and feedback during the course of this investigation. We would also like to thank Peggy D Robertson for her upkeep of the Genome Variation Server and responsiveness in troubleshooting analysis issues.

References

1. P2RX7 purinergic receptor P2 × 7 [Homo sapiens (human)]. 2020, February 3. NCBI. https://www.ncbi.nlm.nih.gov/gene/5027#gene-expression. Accessed 21 February 2020.
2. Benzaquen J, Heeke S, Hreich S, Douguet L, Marquette C, Hofman P, Vouret-Craviari V. Alternative splicing of P2RX7 pre-messenger RNA in health and diseases: Myth or reality? Biomedical Journal. 2019;42(3):141–54. doi:https://doi.org/10.1016/j.bj.2019.05.007.

3. Peng T, Zhong L, Wan Z, Fu W, Sun C. Identification of rs11615992 as a novel regulatory SNP for human P2RX7 by allele-specific expression. Mol Genet Genomics. 2019; 295(1); 23–30; doi:https://doi-org.proxy1.library.jhu.edu/10.1007/s00438-019-01598-0.
4. Fuller SJ, Stokes L, Skarratt KK, Gu BJ, Wiley JS. Genetics of the P2 × 7 receptor and human disease. Purinergic Signalling. 2009;5(2):257–62. doi:https://doi.org/10.1007/s11302-009-9136-4.
5. McQuillin A, Bass NJ, Choudhury K, Puri V, Kosmin M, Lawrence J, Curtis D, Gurling HMD. Case–control studies show that a non-conservativeamino-acid change from a glutamine to arginine in theP2RX7 purinergic receptor protein is associated with both bipolar- and unipolar-affective disorders. Mol Psychiatry. 2009;14(6):614–20. doi:10.1038/mp.2008.6.
6. Auton A, Abecasis G, Altshuler D, et al. A global reference for human genetic variation. Nature. 2015;526:68–74. doi:https://doi.org/10.1038/nature15393.
7. Nickerson D, Rieder M, Carlson C, Yi Q. ldSelect. http://gvs.gs.washington.edu/GVS150/ (2004). University of Washington. Accessed 2 May 2020.
8. NONCODE 2016: an informative and valuable data source of long non-coding RNAs https://doi.org/10.1093/nar/gkv1252
Zhao Y, Li H, Fang S, Kang Y, Wu W, Hao Y, Li Z, Bu D, Sun N, Zhang MQ, Chen R. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nu Ac Res 2016; 44(D1): D203–8; doi:https://doi.org/10.1093/nar/gkv1252.
9. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nuc Ac Res. 2006;34(Database issue):D108–10.