Novel retroduplications discovered in 26 human populations reflect within-population genomic variance and between-population relationships

Yan Zhang 1/9/2015 3:10 PM
Comment [1]: Open to other titles.

Shantao 1/12/2015 12:51 AM
Comment [2]: Suggest we discuss this after finishing the results part

Yan Zhang[1,2]*, Shantao Li[1]*, The 1000 Genomes Project Consortium, Alexej Abyzov[3§], Mark B Gerstein[1,2,4§]

[1]Program in Computational Biology and Bioinformatics, Yale University, BASS 432&437, 266 Whitney Avenue, New Haven, CT 06520, USA

[2]Department of Molecular Biophysics and Biochemistry, School of Medicine, Yale University, 266 Whitney Ave, New Haven, CT 06520, USA

[3]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Harwick 3-12, Mayo Clinic, 200 1st street SW, Rochester, MN 55905, USA

[4]Department of Computer Science, Yale University, 51 Prospect Street, New Haven, CT 06511, USA

*These authors contributed equally to this work

[§]Corresponding authors

Email addresses:

YZ: yan.zhang.yz464@yale.edu

SL: shantao.li@yale.edu

AA: abyzov.alexej@mayo.edu

MBG: mark.gerstein@yale.edu

Abstract

Yan Zhang 1/11/2015 1:08 AM
Comment [3]: "The Abstract of the manuscript should not exceed 250 words and must be structured into separate sections: Background, the context and purpose of the study; Results, the main findings; Conclusions, brief summary and potential implications. Please minimize the use of abbreviations and do not cite references in the abstract."

Background

Text for this section of the abstract…

Results

Text for this section of the abstract…

Conclusions

Text for this section of the abstract…

Keywords

Retroduplication, whole exome sequencing, phylogenetic tree, association, functional impact

Yan Zhang 1/11/2015 1:09 AM
Comment [4]: "Three to ten keywords representing the main content of the article."

Background

Retrotransposons are class I transposable elements. In a retrotransposition event, an element is first transcribed into RNA, and then reverse transcribed back into DNA, which is inserted into a new position in the genome. It has been found that L1 retrotransponsons, the only autonomous mobile elements in human genome, also pick up cellular mRNAs as templates for reverse transcription and insertion (Esnault C et al., 2000; Wei Gilbert N et al., 2001; Mandal PK et al., 2013), creating retroduplications. Although retroduplication forms an exception of the central dogma, it is far more common than we expected before. Three recent studies (Abyzov, Ewing, Schrider) have revealed extensive retroduplication polymorphism in human genomes.

Gene retroduplication represents an important mechanism of gene duplication. Therefore it contributes to the genome evolution and new gene generation in large time scale (Kaessmann H et al., 2009; Ciomborowska et al., 2012; Long M et al., 2013). While some of the retroduplications can become retrogenes with protein coding ability, most of the retroduplications (called processed pseudogenes) are not protein coding genes, since they suffer from the lack of promoter, 5' truncation, mutations, inactive local chromatin environment etc. that hinder the expression of functional protein products. However, the latter also exhibit functional impacts at times. Some of them are able to transcribe and produce protein product (Breyer et al., 2014). In some cases, cellular environment change, such as cancer initiation, can "activate" retroduplications, and both transcription and translation evidence have been observed (The ENCODE Project Consortium 2012; Pei et al., 2012; Sisu et al., 2014). In other cases, transcription products play a role in expression regulation of their parent genes (Sasidharan and Gerstein 2008; Salmena et al. 2011). Two known regulatory mechanisms are RNA inference (Tam et al. 2008; Watanabe et al. 2008; Wen et al. 2011) and retroduplications' transcription products serving as competitive

Yan Zhang 1/10/2015 3:20 PM
Comment [5]: "The Background section should be written in a way that is accessible to researchers without specialist knowledge in that area and must clearly state - and, if helpful, illustrate - the background to the research and its aims. The section should end with a brief statement of what is being reported in the article."

Shantao 1/11/2015 11:06 PM
Comment [6]: Rough working draft below; citations/format to be merged

Yan Zhang 1/21/2015 3:55 PM
Deleted: While

Yan Zhang 1/21/2015 3:59 PM
Deleted: retroduplications

Yan Zhang 1/21/2015 4:01 PM
Deleted: the retroduplications

miRNAs binding targets (Betran E et al., 2004; Poliseno et al. 2010). Recent studies showed such expressed retroduplications result in cancer susceptibility (Breyer et al., 2014). Sometimes retroduplications can have high impacts if they insert into functional regions, of which normal functions are disrupted. Studies have confirmed cases in which germline intragenetic retroduplications result in liver cancer susceptibility (Shukla et al., 2013) and primary immunodeficiency (de Boer M et al., 2014). Besides germline events, a number of researchers have reported massive somatic retroduplications events and their critical roles in tumor development (Solyom et al., 2012; Shukla et al., 2013; Cooke et al., 2014; Tubio et al., 2014; Helman et al., 2014), as well as in neuron development (Richardson et al., 2014; Evrony et al., 2015).

Retroduplications carry several distinctive features: exon-exon junctions, genome locations distant to parent genes, poly-A tail and L1 transposition markers such as target-site duplication (TSD) and human L1 endonuclease preferential cleavage site. In this study, we developed novel methods to exploit these features and performed a comprehensive discovery and analysis of novel retroduplications in 2,535 individuals from 26 populations, without differentiating retrogenes from processed pseudogenes. This largest germline retroduplication polymorphism landscape to date gives us the power to … [[STL: maybe we should fill in this later]]

Yan Zhang 1/21/2015 4:07 PM
Deleted: .

Results and discussion

Yan Zhang 1/10/2015 9:46 PM
Comment [7]: "The Results and discussion may be combined into a single section or presented separately. The Results and discussion sections may also be broken into subsections with short, informative headings."

Novel retroduplications

The 1000 Genomes Project Phase 3 has produced high coverage whole exome sequencing (WXS) data for 2,535 normal individuals from 26 populations {/cite 1000Genomes phase3}. We extracted unmapped reads from the data, and mapped them against exon junction libraries that we built from protein coding exons (see **Calling pipeline** in **Methods**, **Figure 1**). From our pipeline, a total of 15,694 retroduplications have been called from all individuals without merging calls, supported by a total of 63,369 exon junctions, mapped to 740,008 supporting reads. On average, each individual has 6 novel retroduplications identified. NA19318 from LWK (Luhya in Webuye, Kenya) is the only exception that has no novel retroduplication call at all. The novel retroduplications originated from 529 unique parent genes. The retroduplication call set is provided in **Additional file 1**. We further

identified 112 retroduplication insertion sites using extra information from paired end reads (**Additional file 2**). The summary of the call set is shown in **Supplementary Table 1** and **Supplementary Figure 1**.

Novel retroduplications are the retroduplications not annotated in the reference genome. They are not always shared among individuals from the same population. The difference reflects within-population genomic variance. Variance of population-specific novel retroduplications likely occurred in recent human evolution history when the population had diverged from the rest. On the other hand, most novel retroduplication parent genes (XX%) are exclusively identified in a single population, and there are also ones commonly identified in multiple populations (**Supplementary Figure 1B**). We can hypothesize that the common retroduplications were inherited from a common ancestry, while the exclusive retroduplications developed after population divergence. The common and exclusive retroduplications reflect the relationship between different populations, and can be used for assigning populations along human phylogenetic trees.

I think this part can be made stronger by:

1) Providing genotypes for each call (need to include zygocity). Here we can also check for HWE
2) Calculating taggability by SNPs (for calls with detected insertion point)

Any ideas of describing software? Or writing application note about it?

Section about transduction here ?

Can we also explain missing insertion sites by mRNA being used as a template in NHEJ?

http://www.ncbi.nlm.nih.gov/pubmed/?term=Repair+of+DNA+double-strand+breaks+by+templated+nucleotide+sequence+insertions+derived+from+distant+regions+of+the+genome

Phylogenetic tree based on novel retroduplications

Before building phylogenetic trees, we diagnosed the number of novel retroduplications detected in each individual from the same populations. We detected

Yan Zhang 1/10/2015 8:32 PM
Comment [8]: Shantao, please update the number if you need. If you have more summary sentences of insertion sites, write it in this session.

Yan Zhang 1/15/2015 12:04 PM
Comment [9]: Yan will add in a suppl. Fig/table comparing our results and several other studies. Maybe Venn diagrams.

Shantao 1/12/2015 12:17 AM
Comment [10]: Genotyping, Verification in high-cov trios and some data overview will be inserted here later

two outlier individuals, NA11994 in CEU, and NA19042 in LWK, with abnormally high number of novel retroduplication calls (**Supplementary Figure 2**), and they were ruled out from the phylogenetic analysis. The outliers might be due to experimental artifacts or data quality issue.

Taking both within-population genomic variance, in terms of various retroduplication frequencies, and between-population distance into account, we constructed two phylogenetic trees through bootstrap resampling (see **Methods**, and **Figure 2**). One for all 26 populations enrolled in the 1000 Genomes Project [1], and the other for the 17 non-mixed populations (ACB, ASW, CDX, CEU, CHB, CHS, ESN, FIN, GBR, GWD, IBS, JPT, KHV, LWK, MSL, TSI, and YRI). From both phylogenetic trees, we see that the superpopulation groups - Africans, Asians, and Europeans - are clustered with high confidence.

In the all population tree (**Figure 2A**), African superpopulation substructure has AU (approximately unbiased) probability value [2, 3] 0.99. AU is a value between 0 and 1. The higher the AU value is, the more certain the sub-tree structure is. East Asian populations are tightly clustered (AU = 0.78), where all populations in China (CHB, CHS, and CDX) are even tighter clustered (AU = 0.91). Chinese Dai (CDX) is clustered next to Vietnamese (KHV), which might be explained by their geographical closeness. Peruvians (PEL), a mixed population, is clustered with East Asian populations (AU = 0.77). All European populations are clustered in a superpopulation (AU = 0.83), mingled with mixed populations from South Asia and South America. In the non-mixed population tree (**Figure 2B**), African superpoplation (AU = 1), East Asian superpopulation (AU = 0.97), and European superpopulation (AU = 0.94) are more clearly tightly clustered, respectively.

In line with population genetics one can check whether there are signs of population differentiation for different retrodups.

Association between novel retroduplication and gene expression

In each population enrolled in the Geuvadis RNA-sequencing project [4], we investigated the association between novel retroduplication and gene expression on two different time scales (see **Analyze association between retroduplication and gene expression** in **Methods**). The short time scale events (retroduplication variance) happened more recently in human history than long time scale events (change in gene expression level).

Shantao 1/13/2015 10:43 AM
Comment [11]: Justify for separating "small/large time scale"

In the short time scale analysis, we wanted to answer: given the parent genes having differential retroduplication occurrence in a population, is having such novel retroduplications or not associated with parent gene expression? In our within-population tests, we did not see a variable novel duplication significantly associated with its parent gene expression. The global effects of all the retroduplication parent genes are also not significant: Utah residents with Northern and Western European ancestry (p-value 0.579), Finnish (p-value 0.883), British (p-value 0.570), Toscani (p-value 0.912), and Yoruba (p-value 0.274) (**Additional file 3 sheet 1**). It means that novel retroduplication variance is only genomic polymorphism happening in recent human history, and the parent gene expression among individuals in the same population have not differentiated. Note that the statistical power might be limited by the number of novel retroduplications in each population.

In the long time scale analysis, we wanted to answer: are these parent genes highly expressed compared to the background of all the genes? The answer is yes, and the association is significant in all tested populations: Utah residents with Northern and Western European ancestry (p-value 3.23×10^{-12}), Finnish (p-value 1.34×10^{-5}), British (p-value 8.82×10^{-8}), Toscani (p-value 3.78×10^{-10}), and Yoruba (p-value 1.09×10^{-7}) (**Additional file 3 sheet 2**). It means the novel retroduplications came from highly expressed genes. Although examples have been observed that expressed pseudogenes might influence the expression level of parent genes [5, 6], it has been found that processed pseudogenes are rarely transcribed (~4-6% of all processed pseudogenes) in human genome [7]. Thus, parent gene expression level might be a factor influencing retroduplication generation. It is consistent with our knowledge that the more mRNAs a gene has made, the higher probability that it will be converted into complementary DNA and inserted back into the genome. {\cite publication?}.

Differential expression of retroduplication parent genes

Differential expression of retroduplication parent genes between populations

[[Yan: coming soon.]]

Tissue-specific expression analysis[[Differential expression of retroduplication parent genes between tissues]]

As we have shown genes highly expressed in general are prone to retroduplication, we are interested in whether such expression patterns bias toward certain tissues. Indeed,

Yan Zhang 1/21/2015 4:27 PM
Comment [12]: Extra eQTL and between-population tests will be performed.

Alexej Abyzov 1/21/2015 7:37 PM
Comment [13]: What is global effect?

Yan Zhang 1/21/2015 4:22 PM
Deleted: we did not see significant evidence supporting the association within each population:

Shantao 1/11/2015 11:14 PM
Comment [14]: Could be due to purifying selection

Alexej Abyzov 1/21/2015 7:40 PM
Comment [15]: This can also mean that rdup has no effect of expression of the parent gene.

Yan Zhang 1/15/2015 12:36 PM
Comment [16]: I am thinking about merging Alex's (and others') novel retroduplication sets, and perform the analysis again. This results using only our retrodups can be put into supplement.

Alexej Abyzov 1/22/2015 2:22 PM
Comment [17]: One can also check eQTL of genes near insertion points.

Yan Zhang 1/21/2015 4:56 PM
Deleted: It is consistent with our knowledge that genes with high expression level tend to have more copies (including retrogenes and processed pseudogenes) in the genome {\cite publication?}.

Alexej Abyzov 1/22/2015 2:26 PM
Comment [18]: How is it different from the analysis above?

to be carried in germline, retroduplications should occur only in germline cells and early embryonic development. Several previous studies have shown some known retroduplicated genes are expressed in early embryonic cells (Booth and Holland, 2004 [[PMID: 15233988]]; Elliman et al., 2006 [[PMID: 16291741]]; Pain et al., 2005 [[PMID: 15640145]]) and germline linkage cells (Pain et al., 2005 [[PMID: 15640145]]; Malki et al., 2014 [[PMID: 24882376]]).

Using tissue specific expression dataset obtained from TiGER (Liu et al., 2008) and Expression Atlas (Kapushesky, M. et al. 2012; Petryszak, R. et al. 2013).

[[STL: work in progress; simple aggregation/heat map does not show anything. Will do more rigorous stat tests and maybe use the GTex dataset]]

[[STL: some working ideas below]]

Explore association between retroduplications and methylation patterns in germline cells and early embryonic cells, sperm and oocytes?

Explore association of insertion points and open chromatin markers

Would be good to compare with TEI from 1KG.

Overlap retroduplications with somatic events

[[Yan: I will check Peter Campbell's paper for this section.]]

Function impact of novel retrodplications

Functional enrichment analysis

We performed functional enrichment analysis for the 529 unique parent genes. The enriched terms and significance scores are listed in **Table 1**. Terms related to ribosome/structural molecule activity, intracellular organelle lumen/nucleoplasm, and protein complex assembly are among the most enriched. This observation is in accordance with previous study [8] indicating retrotransposition is coupled with cell division. Intracellular lumen is related to tubular structures inside cells, and is involved in membrane formation, which is a crucial step in cell formation/division. Ribosome activity and protein complex assembly are also contributing to new protein production for a new cell. [[Yan: might update.]]

Overlap between genomic elements and retroduplication insertion sites

[[Yan: coming soon.]]

Conclusions

We conducted outstanding analysis and got superb results. Without our study science would stop. Text for this section.

Yan Zhang 1/10/2015 5:06 PM
Comment [19]: "This should state clearly the main conclusions of the research and give a clear explanation of their importance and relevance. Summary illustrations may be included."

Methods

Yan Zhang 1/6/2015 10:59 PM
Comment [20]: Use present tense in Methods.

Data resources

Whole exome sequencing (WXS) and whole genome sequencing (WGS) data of 2,535 individuals from 26 populations are provided by the 1000 Genomes Project Phase 3 {\cite 1000G phase 3}. Population description can be found at http://www.1000genomes.org/category/frequently-asked-questions/population. Protein-coding gene expression data (Peer-factor normalized RPKM) is obtained from the Geuvadis RNA-sequencing project [4], which generated RNA sequencing data from lymphoblastoid cell lines of 462 individuals from 5 populations (CEU, FIN, GBR, TSI and YRI) enrolled in the 1000 Genomes Project. We use human reference genome build 37 [9], and GENCODE v19 human genome annotation [10] in the study.

Calling pipeline

The calling pipeline is customized for generating retroduplication calls from high-coverage exome sequencing data. It is adapted and adjusted from our previous pipeline developed for low-coverage whole genome sequencing data [8]. A simplified flowchart of the current pipeline is shown in **Figure 1**.

Build and index true and decoy exon junction libraries

For calling retroduplications from whole exome sequencing data, we need to build exon junction libraries from annotated protein coding exons. The true exon junction library is built by joining pairs of protein coding exon segments within the same genes, while maintaining exons' order on the strand. Exon segments of length 100 bases adjacent to the joining splice sites are combined (**Supplementary Figure 3**). We also build five decoy exon junction libraries for the purpose of controlling false call rate (FCR). The decoy exon junction libraries contain fake exon junctions, in which exon annotations are shifted by *e* base(s) on both sides (i.e. start location + *e*, end location - *e*). *e* is taken as 1, 2, 3, 6, and 12 for each decoy exon library,

Yan Zhang 1/7/2015 4:11 PM
Comment [21]: Alex, this is directed read out from the code. This means the exons being joint are shifted apart by 2e.

respectively. Subsequently, all the junction libraries are indexed using BWA-0.7.7 [11] for the purpose of speeding up following read alignment.

Generate unmapped read alignments

Unmapped reads can be utilized for calling novel retroduplications that are not annotated in the reference genome. We use SAMtools [12] to extract unmapped reads from exome bam files, then use BWA-0.7.7 to align the unmapped reads to all of true and decoy exon junction libraries (**Supplementary Figure 3**). BWA default parameters of allowed mismatches are used in the alignment. d_1 and d_2 are the number of bases that the read maps to either exon segment. $\min(d_1, d_2) \geq d$ is required for a newly mapped read to be reported from our pipeline. d is a parameter automatically tuned in the range [0.005, 0.05], ensuring the most number of calls from the true exon junction library while satisfying the false call rate (FCR) ≤ 0.05. FCR is an empirical measure of false discovery rate, and it is calculated as the maximum number of mapped reads among all decoy exon junction libraries divided by the number of mapped reads in the true exon junction library.

Generate novel retroduplication calls

Multiple "previously unmapped" reads might be mapped to the same exon junction, supporting the existence of the exon junction. Also, multiple exon junctions with supporting evidence might come from the same gene. We report the gene having novel retroduplications, when it has at least 2 supporting exon junctions with newly mapped reads. The genes (also called *parent genes*) with novel retroduplications are called for each person.

Detect retroduplication insertion sites

We search for discordant sequencing pairs (with a minimum quality score of 15) with one read correctly mapped to parent gene and the other read mapped to a different chromosome or at least 1kb far away from the gene. To avoid mismapping problem, we also exclude the case when the other read mapped within 1kb to Gencode level 2 pseudogene that is originated from the gene (and has a sequence identity of at least 97%).

Pairs with proper orientations are clustered using average linkage clustering. It can be showed that this linkage criterion is not affected by the local coverage. Assuming uniform distribution of reads, it can be shown mathematically that the expected distance is {2(IS - RL) + 1} / 3, where the IS is the insertion size and RL is the read length [[STL: Supplemental?]]. As the insertion size in most cases is 200-400bp, we choose 500bp as the cut-off for average linkage distance to stop clustering.

Yan Zhang 1/6/2015 11:53 PM
Comment [22]: Alex, is it accurate for me to say so?

Yan Zhang 1/8/2015 7:34 PM
Comment [23]: Extract only one unmapped read, if one map but the other does not.

Yan Zhang 1/8/2015 8:08 PM
Comment [24]: Will double check scenarios.

Shantao 1/11/2015 11:49 PM
Comment [25]: Need further investigation on this.

This cut-off not only takes the deviations of insertion size into consideration, but also allows sufficient space for TSD. A valid insertion point must have at least two reads on one side (i.e. stand). [[STL: Some schemes and updates for rescuing one-side cases]]

Genotype novel retroduplications

Explore association between retroduplications and methylation patterns in germline cells and early embryonic cells

Explore association of insertion points and open chromatin markers

Build population phylogenetic trees based on novel retroduplication calls

Rule out outlier individuals

In each population, we sort the individuals based on the number of unique parent genes having novel retroduplications (**Supplementary Figure 2**). The individuals with exceptionally high number of parent gene identifications are ruled out from the phylogenetic analysis.

Generate retroduplication frequency matrix

There are parent genes called in common among multiple populations, or called exclusively in one population. On the other hand, within a population, the parent gene calls are not appearing at the same frequency. This information can be used for measuring distance between populations, while taking into account different retroduplication frequencies. We define a retroduplication frequency matrix, from which distance measures can be calculated.

Suppose there are N populations, and M unique parent genes are identified in these populations. The retroduplication frequency matrix A is defined as an $M \times N$ matrix, with each element $A_{m,n}$ ($m=1,2,...,M$; $n=1,2,...,N$) being a value in [0, 1], representing the percentage of individuals in population n having this unique parent gene m called.

Bootstrap phylogenetic trees

We use Manhattan distance as the distance measure between each pair of populations (i.e. Manhattan distance between two columns in A). Agglomerative method "average" is used in hierarchical clustering for generating each tree. 1000 bootstrap replications are performed and the uncertainty is assessed using Pvclust [2]. The

reported AU (approximately unbiased) probability values [2, 3] are used to indicate the certainty of sub-tree structures generated from multi-scale bootstrap resampling [13–15]. The higher the AU probability value, the more confident the substructure is.

Analyze association between retroduplication and gene expression

We utilize our retroduplication call set and the Geuvadis gene expression data (Peer-factor normalized RPKM) [4] to analyze the association between retroduplication occurrence and gene expression. Matching data of the individuals enrolled in both the 1000 Genomes Project and the Geuvadis project are used. The association can be investigated on two different time scales. (1) The "short" time scale analysis focuses on the variance of novel retroduplications within the parent gene set for each population. A parent gene has novel retroduplication(s) in at least one individual in the population. (2) The "long" time scale analysis inspects all the genes, and tests whether retroduplication parent genes tend to be highly expressed. The short time scale events (retroduplication variance) happened more recently in human history than long time scale events (change in gene expression level). The association tests are performed for each population, respectively, in order to rule out the confounding by population stratification.

Shantao 1/13/2015 10:52 AM
Comment [26]: Again, don't think so

Analyze association in short time scale

For a certain population, we perform the association test within the set of retroduplication parent genes: test whether having novel retroduplication(s) or not is associated with the parent gene's expression level.

First, differential expression of each parent gene is tested between the group of individuals that have novel retroduplications of this gene and the group of individuals that do not. Two-sided Wilcoxon rank sum test is used. p-values are adjusted by Benjamini-Hochberg procedure [16]. A gene is reported to be differentially expressed in the parent gene set if its adjusted p-value is less than 0.05. Furthermore, the global differential expression of all the parent gene set is tested using Fisher's combined probability test [17] on unadjusted p-values. It can test the combined effect of multiple parent genes, whose individual effect is not necessarily strong. If the combined p-value is less than 0.05, we can conclude that the association between retroduplication variance and parent gene expression is significant.

Shantao 1/11/2015 11:19 PM
Deleted: are

To re-confirm the result, we also perform two-sided Wilcoxon signed rank test. For each gene, medium expressions of both groups (having the novel

retroduplication or not) are paired. The test result is consistent with that of the Fisher's method.

Analyze association in long time scale

For a certain population, we test whether the retroduplication parent genes are highly expressed among all the genes measured in the Geuvadis data set. We take medium expression value over all individuals for each gene as the representative expression value. One-tailed empirical p-value is calculated comparing the expression value of each parent gene versus the null distribution of expression values of all genes. It indicates the significance of each retroduplication parent gene having high expression value among all genes. Fisher's combined probability test is performed on the empirical p-values. If the combined p-value is less than 0.05, that means in general the parent genes are significantly highly expressed among all genes.

Analyze differential expressions of retroduplication parent genes

Differential expression of retroduplication parent genes between populations

Tissue-specific expression analysis **[[Differential expression of retroduplication parent genes between tissues]]**

[[Shantao, insert your text here]]

Investigate functional impact of novel retroduplications

Functional enrichment analysis

We use DAVID [18] to annotate functional terms (including GO terms [19]) for retroduplication parent genes, and survey functional term enrichment.

Overlap genomic elements and retroduplication insertion sites

[[Yan: coming soon.]]

Abbreviations

WXS: whole exome sequencing; WGS: whole genome sequencing; FCR: false call rate; AU: approximately unbiased.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AA, YZ and MBG conceived the study. YZ and SL performed the analyses. YZ, SL, AA and MBG wrote the manuscript. YZ generated retroduplication calls based on exon junction libraries, and carried out phylogenetic analysis, association analysis, and functional enrichment analysis. SL carried out retroduplication insertion site detection, and tissue-specific expression analysis. The 1000 Genomes Project Consortium Phase 3 provides DNA sequencing data. AA and MBG co-directed the work.

Acknowledgements

The work was supported by XXX funding. The authors would like to thank XXX for useful discussion and proofreading the manuscript.

References

1. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**:1061–1073.

2. Suzuki R, Shimodaira H: **Pvclust: an R package for assessing the uncertainty in hierarchical clustering.** *Bioinformatics* 2006, **22**:1540–2.

3. Shimodaira H, Hasegawa M: **CONSEL: for assessing the confidence of phylogenetic tree selection**. *Bioinformatics* 2001, **17**:1246–1247.

4. Lappalainen T, Sammeth M, Friedl\``ander MR, ’t Hoen PAC, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Alml’'of J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HPJ, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, et al.: **Transcriptome and genome sequencing uncovers functional variation in humans**. *Nature* 2013, **501**:506–511.

5. Chiefari E, Iiritano S, Paonessa F, Le Pera I, Arcidiacono B, Filocamo M, Foti D, Liebhaber SA, Brunetti A: **Pseudogene-mediated posttranscriptional silencing of HMGA1 can result in insulin resistance and type 2 diabetes.** *Nat Commun* 2010, **1**:40.

6. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP: **A coding-independent function of gene and pseudogene mRNAs regulates tumour biology.** *Nature* 2010, **465**:1033–8.

7. Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M: **Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability**. *Nucleic Acids Res* 2005, **33**:2374–2383.

8. Abyzov A, Iskow R, Gokcumen O, Radke DW, Balasubramanian S, Pei B, Habegger L, Lee C, Gerstein M: **Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division.** *Genome Res* 2013.

9. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, et al.: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860–921.

10. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, et al.: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res* 2012, **22**:1760–74.

11. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754–60.

12. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078–9.

13. Efron B, Halloran E, Holmes S: **Bootstrap confidence levels for phylogenetic trees**. *Proc Natl Acad Sci U S A* 1996, **93**:13429–13434.

14. Shimodaira H: **An approximately unbiased test of phylogenetic tree selection.** *Syst Biol* 2002, **51**:492–508.

15. Shimodaira H: **Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling**. *Ann Stat* 2004, **32**:2616–2641.

16. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *J R Stat Soc Ser B* 1995, **57**:289–300.

17. Fisher RA: *Statistical Methods for Research Workers*. Oliver and Boyd; 1925(no 5):xv, 356 p.

18. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4**:44–57.

19. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25–9.

Figures

Figure 1 - Simplified flowchart of the retroduplication calling pipeline.

Figure 2 - Phylogenetic trees built based on novel retroduplications

A – The phylogenetic tree of all 26 populations enrolled in the 1000 Genome Project. B – The phylogenetic tree of 17 non-mixed population. Red number: AU (approximately unbiased) probability value. Green number: BP value, i.e. the

frequency of the cluster appearing in bootstrap replicates. In the rectangulars, we highlight clusters with AU ≥ 0.95. Bootstrap resampling was performed 1000 times for generating the trees shown in A and B.

Tables

Table 1 – Functional enrichment analysis results.
[[Yan: Will be tidied and updated.]]