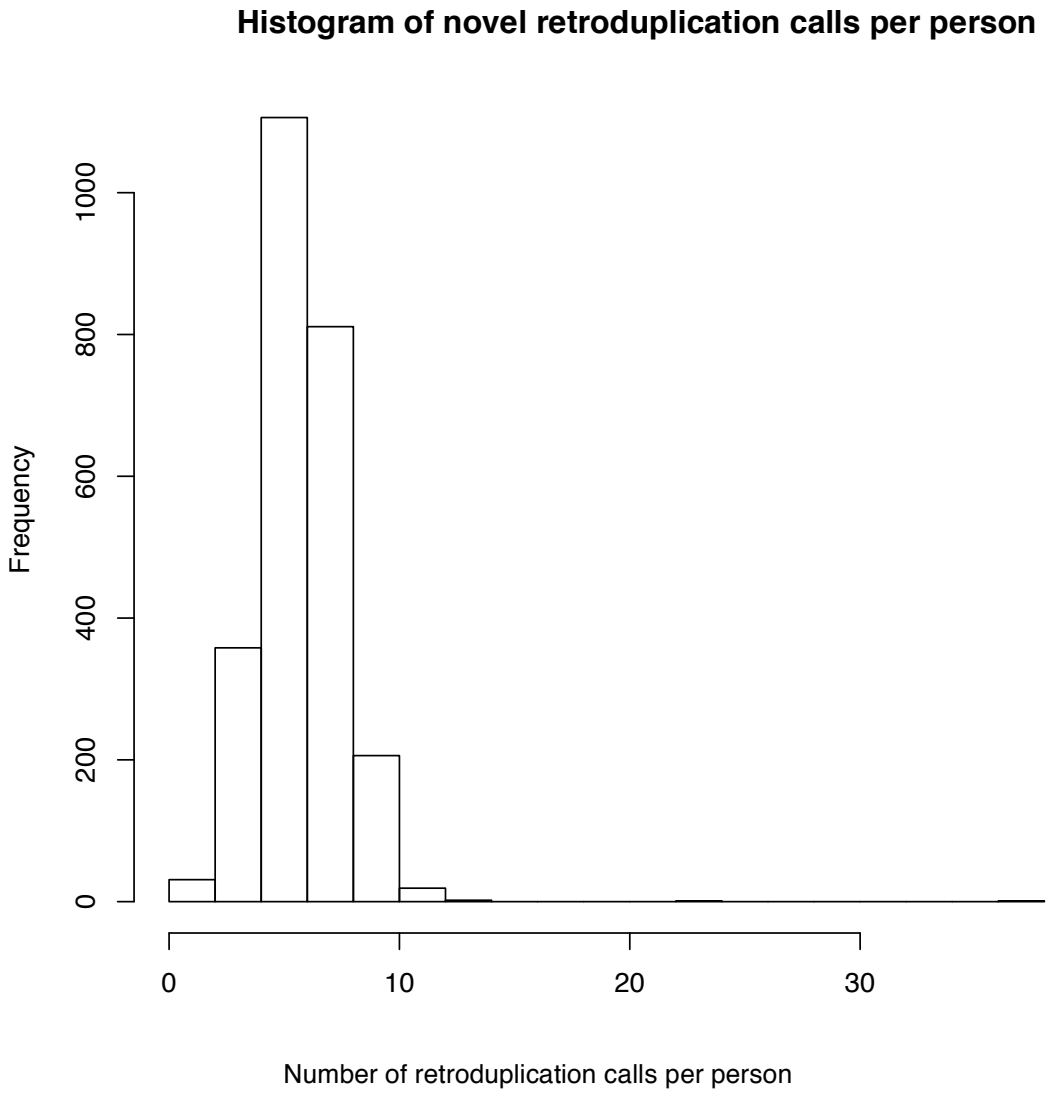
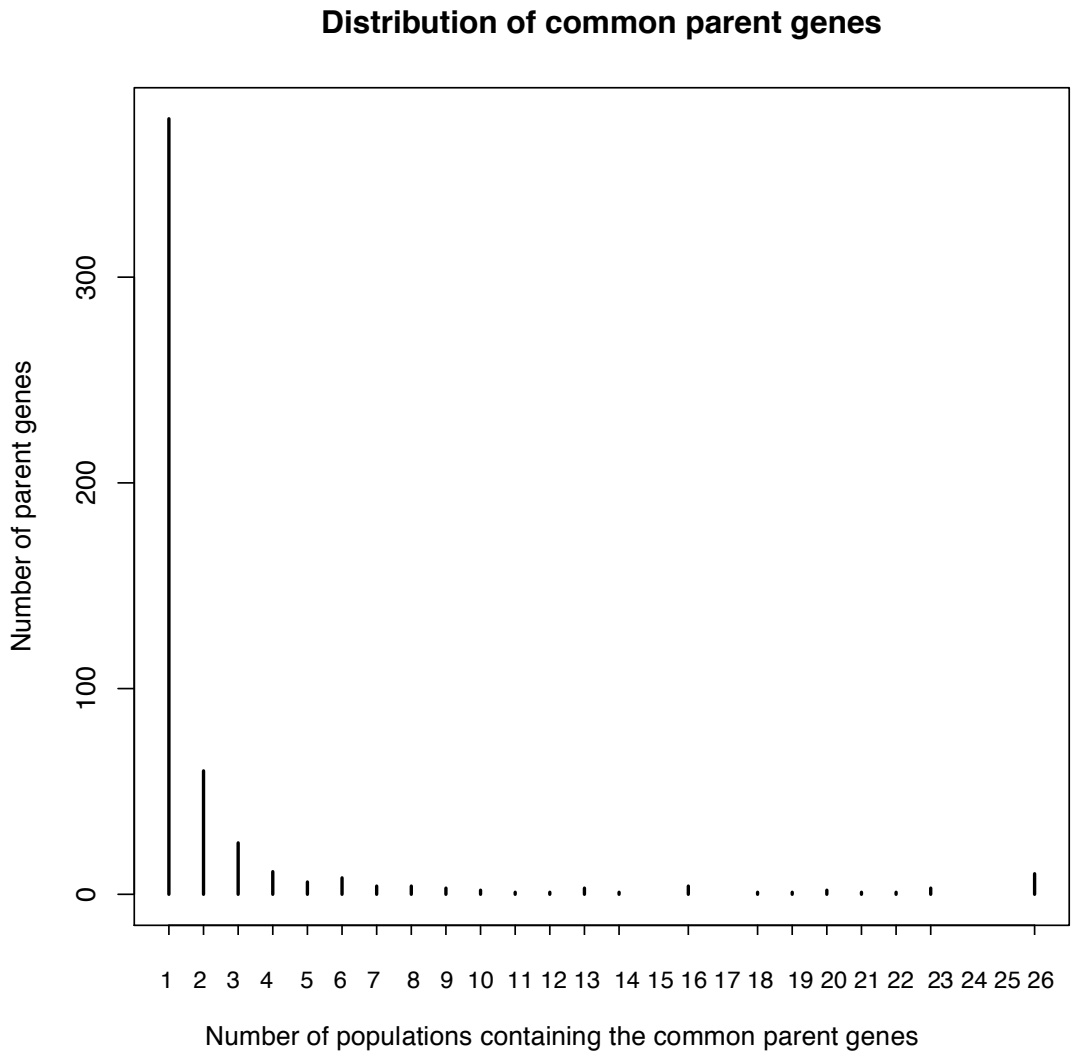
A)

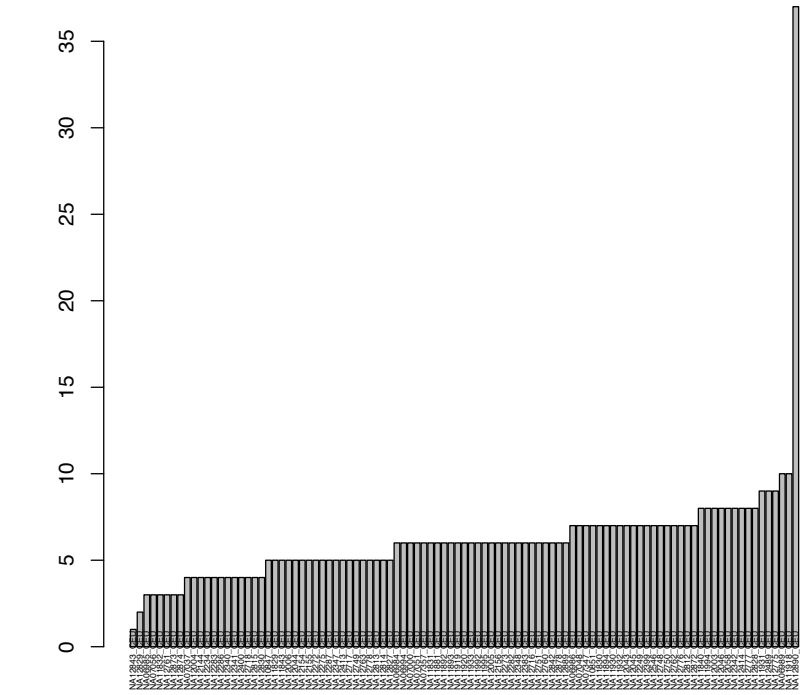

Histogram of novel retroduplication calls per person
Frequency
0
200
400
600
800
1000
0
10
20
30
Number of retroduplication calls per person

B)

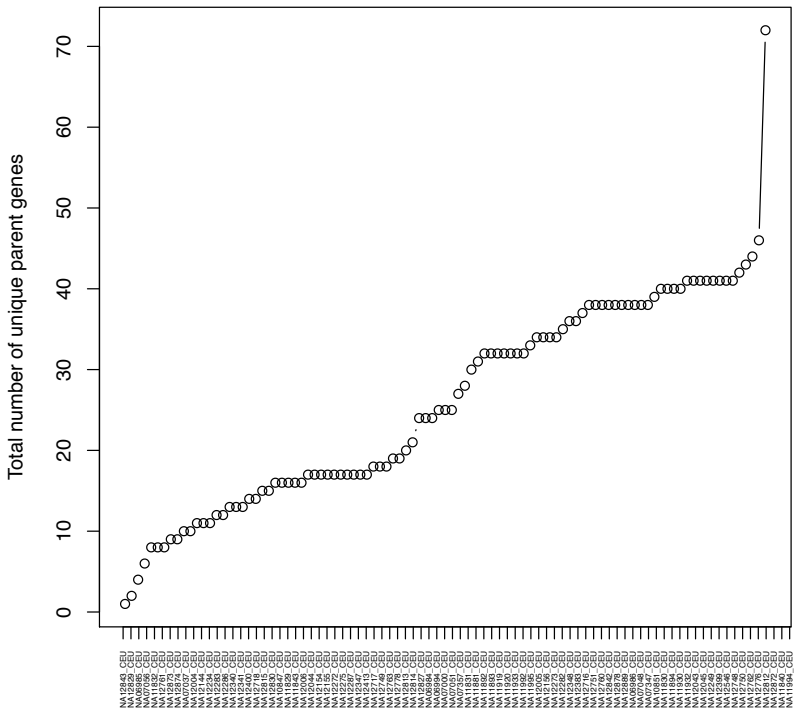


Supplementary Figure 1 – Summary statistics of the retroduplication call set generated based exon junction libraries. A – Histogram of novel retroduplication calls per person. B – Distribution of common parent genes among populations.

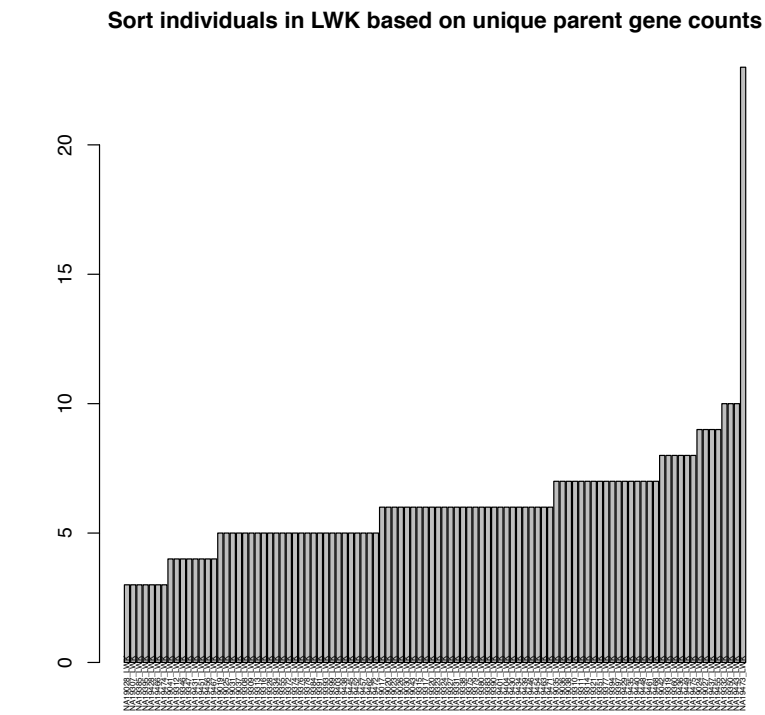
A)


Sort individuals in CEU based on unique parent gene counts
0
5
10
15
20
25
30
35

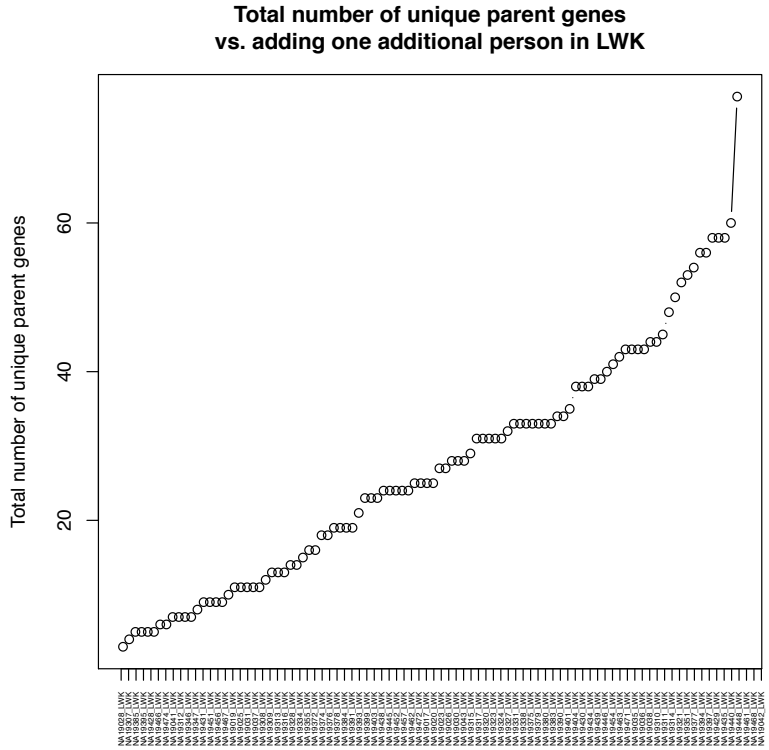
B)


Total number of unique parent genes
vs. adding one additional person in CEU
Total number of unique parent genes
0
10
20
30
40
50
60
70

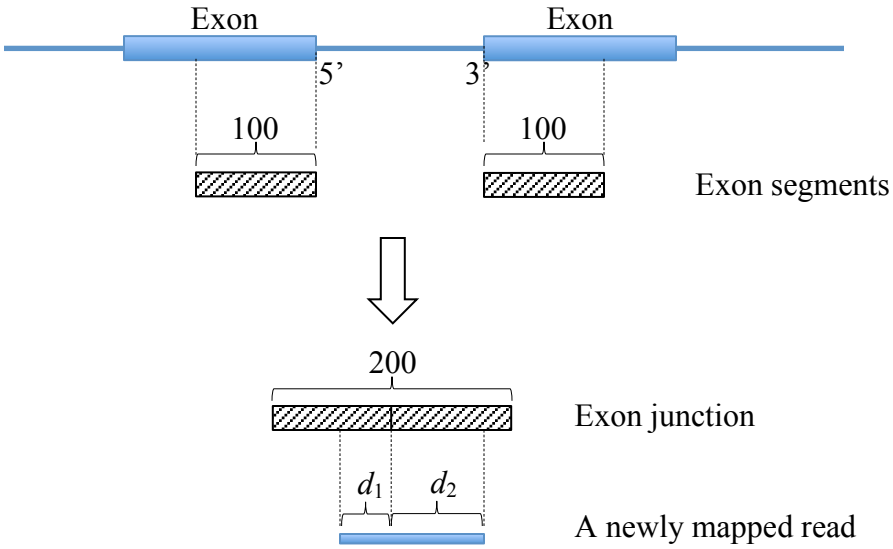
C)


Sort individuals in LWK based on unique parent gene counts
0
5
10
15
20

D)


Total number of unique parent genes
vs. adding one additional person in LWK
Total number of unique parent genes
20
40
60

Supplementary Figure 2 – Diagnosis plots for detection of outlier individuals that have exceptionally high number of retroduplication calls. In each population, we sort the indivuals based on the number of unique parent genes having novel retroduplications. We have detected two individuals with exceptionally high number of retroduplication calls: NA11994 in CEU, and NA19042 in LWK. Diagnosis plots of CEU (panel A, B) and LWK (panel C, D) are shown here. Individuals in each population are sorted by increasing number of retroduplication calls.



Supplementary Figure 3 – Generating exon junctions, and mapping reads to an exon junction. Minimum distance allowed between two exons being joined is 70 bases. Two exon segments adjacent to the joining splice site are extrated, and combined to form an exon junction of length 200 bases. Unmapped reads are mapped against exon junctions. 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. d is an automatically tuned parameter in our pipeline, ensuring the most number of calls while satisfying the false call rate (FCR) ≤ 0.05.

Supplementary Table 1 – Update of Phase 3 retroduplication identifications from Phase 1. In our previous study (Abyzov et al., 2013), we identified retroduplications from the 1000 Genomes Project Phase 1 low coverage whole genome sequencing data. In the new study, we identified retroduplications from Phase 3 high coverage exome sequencing data. Update of this study is summarized in the table.

	Previous study (Abyzov et al., 2013)	New study update
Detection of novel retroduplications using exon junction libraries		
Data	Low coverage whole genome sequencing (WGS)	ILLUMINA high coverage exome sequencing (WXS)
Populations	986 WGS of 14 populations + 2 WGS trios (YRI trio, CEU trio)	2,535 WXS of 26 populations
Number of retroduplication calls per person	On average 6-10	Median: 6, IQR: 2
Number of unique genes with novel exon junctions	147	529
Outlier individuals	6 JPT individuals	1 CEU individual and 1 LWK individual
Insertion site detection using mapping of paired end reads		
Data	968 WGS	2,535 WXS
Number of insertion sites detected	36 insertion points, mapped to 36 unique genes	112 insertion points, mapped to XXX unique genes.

Yan Zhang 1/10/2015 5:48 PM
Comment [1]: Shantao will update these numbers.