Supplementary Text

Principal Components Analysis (PCA):
We performed PCA on each of the four different datasets along with European Americans (CEU), Han Chinese (CHB), and West Africans (YRI), and found that the Siddi in the Human Origins and Affymetrix 6.0 datasets are strong outliers as previously reported (Supplementary Figure 7)[4,13,14]. We next removed YRI, Siddi and indigenous Andamanese (another known outlier) from the datasets and repeated PCA (Figure 1b; Supplementary Figure 6). Similar to past studies, the PCA documents three broad groupings[4,8,11]. First, almost all South Asian groups speaking Indo-European and Dravidian languages lie along the "Indian Cline," consistent with mixture from two highly divergent ancestral populations (Ancestral North Indian (ANI) ancestry related to Europeans, Central Asians, and Near Easterners, and Ancestral South Indian (ASI) ancestry that is as different from ANI as Europeans and East Asians are from each other)[4]. The second major cluster includes groups that speak Austroasiatic languages, as well as some non-Austroasiatic speaking groups that have similar ancestry possibly due to gene flow with Austroasiatic speaking neighbors or due to a history of language shift. This set of groups cluster together near the ASI end of the Indian cline, likely reflecting a large proportion of ASI-related ancestry as well as a distinct ancestry that is related to that in East Asians. Lastly, there are a set of groups with very different population histories that nonetheless all have genetic affinity to East Asian groups. These groups include the Austroasiatic speaking Khasi and Nicobarese, the Tibeto-Burman speaking groups, and several groups with evidence of East Asian mixture including Bengali. We confirmed the evidence of East Asian related mixture in some groups by observing significantly negative f_3(Test; Mala, Chinese) statistics[7] (Supplementary Table 6).

Two Additional Methods to Measure the Strength of Founder Events:
In addition to the IBD-based method for measuring the strength of founder events, we used two other methods that did not require phasing or IBD detection. First, we computed F_{ST} between each group and the closest cluster of groups with similar ancestry sources. Second, for groups on the Indian Cline we fit a model of population history using *qpGraph*[7] and measured the founder event as the group-specific drift after admixture (Supplementary Figure 5 and Supplementary Table 5). These two analyses measure genetic drift, which should be directly correlated with founder event strengths. We found that the results of both methods were highly correlated to that of the IBD-based method for all cases where a comparison was possible. The groups where this was possible consisted of all groups on the Indian cline for the *qpGraph* analyses and both groups on the Indian cline and groups with ancestry similar to Austroasiatic speakers for the F_{ST} analyses (Pearson correlation r=0.82-0.98) (Supplementary Table 3).

Gene Mapping in Two Additional Sets of Patients:
In addition to the 6 PPD patients with Cys78Tyr mutations mentioned in the main text, we genotyped 12 other patients, 6 with PPD and 6 with mucopolysaccharidosis

type IVA (MPS), a disease caused by mutations in the gene *GALNS*[17]. These 12 patients were primarily from consanguineous marriages. We showed that while they could be mapped using genome-wide homozygosity mapping (Supplementary Figures 3b and 3d)[18,19], they did not have substantial IBD across families at the disease mutation site. (If all PPD patients were mapped together, as would likely be the case if the mutations were not known beforehand, the mutation locus would still be found easily (Supplementary Figure 3c)). Moreover, the mutation site haplotypes were smaller than the PPD Cys78Tyr haplotypes (Supplementary Figures 4b and 4d). This suggests that these 2 mutations are at high frequency due to older founder events than the one that occurred for the PPD Cys78Tyr mutations, which could explain why they were not discovered by IBD (which is most sensitive for young founder events within the last dozens of generations) and also why they are present primarily in individuals descending from consanguineous marriages (because they may be sufficiently rare that they do not come together at an appreciable rate except in the context of a consanguineous marriage). Taken together, this demonstrates the importance of the relatively young founder events found in this study (detectable by IBD), which likely lead to an increased burden for recessive diseases even in non-consanguineous contexts.

Total Number of Individuals	3	4	5	6	7	30
Times First Cousin Was Removed	100	100	100	100	100	100
Times Second Cousin was Removed	100	100	100	100	100	100

Supplementary Table 1. 100 simulations to determine sensitivity of relatedness filtering algorithm. Results of performing the relatedness filtering algorithm with 100 simulations of both first and second cousins. The filtering algorithm was able to pick up and remove both the first and second cousin all 100 times in all different sample sizes (total number of individuals refers to the total including the first or second cousins).

IBD Thresholds	Number of Individuals	Times IBD Score of pop1 is greater than that of pop2	Times IBD Score of pop1 is significantly greater than that of pop2	Times IBD Score of pop3 is greater than that of pop2	Times IBD Score of pop3 is significantly greater than that of pop2	Times IBD Score of pop2 is significantly greater than that of pop3
3-20cM	5	100	98	0	0	5
3-20cM	4	100	79	0	0	20
3-20cM	3	98	59	2	0	42
3-20cM	2	87	18	11	0	46
>3 cM	5	99	95	1	0	15
>3 cM	4	99	73	3	0	23
>3 cM	3	98	42	4	0	37
>3 cM	2	83	10	10	0	49

Supplementary Table 2. 100 simulations to determine sensitivity and specificity of IBD analyses. Results of performing IBD analyses with 100 simulations of groups of different bottleneck strengths (see Online Methods for details of the simulations and IBD score calculations). pop2=group with bottleneck approximately the strength of that of Finns based on having a similar IBD score; pop1=group with bottleneck twice as large as that of pop2; pop3=group with bottleneck half as large as that of pop2. This indicates that in groups with twice as strong of a bottleneck as Finns, the sensitivity of the IBD analyses is high with only 4-5 individuals, and in all cases, groups with bottlenecks half the strength of that of Finns never have a significantly higher IBD score than that of Finns. Based on this analysis, we aimed to genotype 5 individuals per group for the new genotyping reported in this study. The IBD thresholds show that adding the additional step of removing IBD segments above 20cM improved the sensitivity of the analyses. There is less sensitivity for determining whether a group has significantly smaller IBD score than that of Finns as shown by the last column.

Genotyping Platform	Comparison Group	Correlation (r)
Human Origins	IBD Score and F_{ST} Score	0.815
Human Origins	IBD Score and Group Specific Drift	0.860
Human Origins	F_{ST} Score and Group Specific Drift	0.961
Affymetrix 6.0	IBD Score and F_{ST} Score	0.856
Affymetrix 6.0	IBD Score and Group Specific Drift	0.913
Affymetrix 6.0	F_{ST} Score and Group Specific Drift	0.965
Illumina	IBD Score and F_{ST} Score	0.975
Illumina	IBD Score and Group Specific Drift	0.937
Illumina	F_{ST} Score and Group Specific Drift	0.960
Illumina_Omni	IBD Score and F_{ST} Score	0.953
Illumina_Omni	IBD Score and Group Specific Drift	0.954
Illumina_Omni	F_{ST} Score and Group Specific Drift	0.964

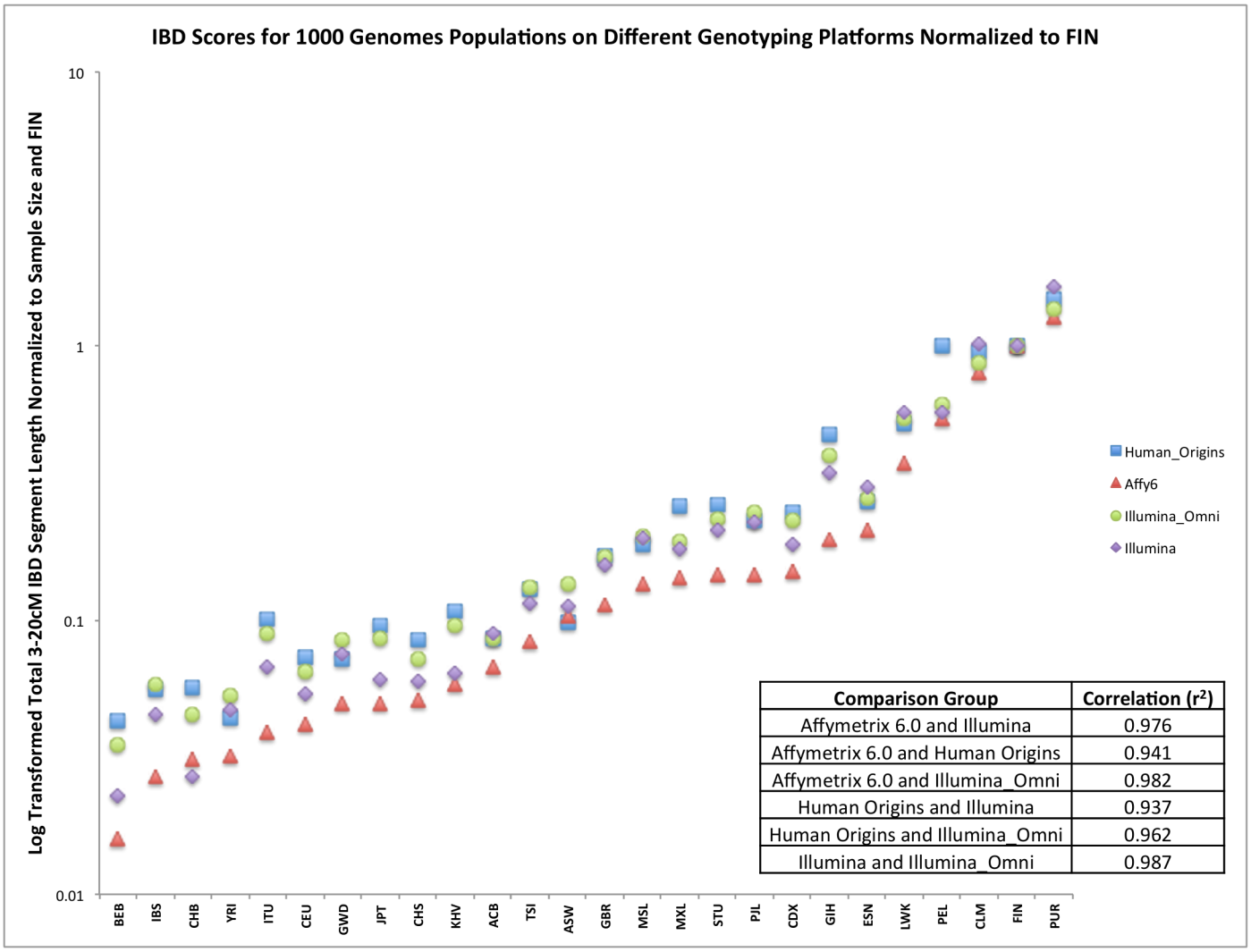
Supplementary Table 3. Measurements of founder event strength. For each group on the Indian Cline, we computed three measures of founder event strength: IBD Score, F_{ST} score, and model-based group-specific drift using *qpGraph* (Online Methods). The three measurements are highly correlated for all cases where a comparison was possible, meaning groups on the Indian cline for the *qpGraph* analyses and both groups on the Indian cline and groups with ancestry similar to Austroasiatic speakers for the F_{ST} analyses.

Location	Linguistic Affiliation	Group	Groups with High Shared IBD
Gujarat	IndoEuropean	Muslim_Jat	Balochi, Muslim_Jat
Pakistan	IndoEuropean	Balochi	Balochi, Brahui, Makrani, Muslim_Jat
Pakistan	Dravidian	Brahui	Balochi, Brahui, Makrani
Pakistan	IndoEuropean	Makrani	Balochi, Brahui, Makrani
Chhattisgarh	IndoEuropean	Ghasia	Satnami
Chhattisgarh	IndoEuropean	Satnami	Gamit, Ghasia, Satnami
Gujarat	IndoEuropean	Gamit	Gamit, Satnami
Maharashtra	IndoEuropean	Warli	Kathodi, Warli
Gujarat	IndoEuropean	Kathodi	Kathodi, Kolcha, Kotwalia, Warli
Gujarat	IndoEuropean	Kolcha	Kathodi, Kolcha, Kotwalia
Gujarat	IndoEuropean	Kotwalia	Kathodi, Kolcha, Kotwalia
Gujarat	IndoEuropean	Patel	Baniya, GujaratiD, Patel
Haryana	IndoEuropean	Baniya	GujaratiD, Jain, Patel
US	IndoEuropean	GujaratiD	Baniya, Patel
Punjab	IndoEuropean	Sikh_Jatt	Punjabi, Scheduled_Caste_Haryana
Punjab	IndoEuropean	Punjabi	Scheduled_Caste_Haryana, Sikh_Jatt
Haryana	IndoEuropean	Scheduled_Caste_Haryana	Punjabi, Scheduled_Caste_Haryana, Scheduled_Caste_Uttrakhand, Sikh_Jatt
Uttar Pradesh	IndoEuropean	Pal	Pal, Tharu_Uttrakhand
Uttarakhand	IndoEuropean	Tharu_Uttrakhand	Pal
Uttarakhand	IndoEuropean	Syon	Syon, Wan
Uttarakhand	IndoEuropean	Wan	Syon, Wan
Goa	IndoEuropean	Brahmin_Catholic_Goa	Brahmin_Catholic_Goa, Brahmin_Catholic_Kumta, Brahmin_Catholic_Mangalore
Karnataka_(Kumta)	IndoEuropean	Brahmin_Catholic_Kumta	Brahmin_Catholic_Goa, Brahmin_Catholic_Kumta, Brahmin_Catholic_Mangalore
Karnataka_(Mangalore)	IndoEuropean	Brahmin_Catholic_Mangalore	Brahmin_Catholic_Goa, Brahmin_Catholic_Kumta, Brahmin_Catholic_Mangalore
Uttarakhand	IndoEuropean	Brahmin_Uttrakhand	Brahmin_Nepal, Brahmin_Uttrakhand
Nepal	IndoEuropean	Brahmin_Nepal	Brahmin_Nepal, Brahmin_Uttrakhand
Nagaland	TibetoBurman	Chakehshanega	Chakehshanega, Nagaseema, Poumainaga
Nagaland	TibetoBurman	Nagaseema	Chakehshanega, Nagaseema, Poumainaga
Manipur	TibetoBurman	Poumainaga	Chakehshanega, Nagaseema, Poumainaga
Karnataka	Dravidian	Hakki_Pikki	Hakki_Pikki, Malaikuravar, Narikuravar
Tamil Nadu	Dravidian	Malaikuravar	Hakki_Pikki, Malaikuravar, Narikuravar
Tamil Nadu	Dravidian	Narikuravar	Hakki_Pikki, Malaikuravar, Narikuravar
Tamil Nadu	Dravidian	Indumalayali	Indumalayali, Irula
Tamil Nadu	Dravidian	Irula	Indumalayali, Irula
Tamil Nadu	Dravidian	Palliyar	Palliyar, Pulliyar
Tamil Nadu	Dravidian	Pulliyar	Palliyar, Pulliyar
Andhra Pradesh	Dravidian	Oddari	Korava, Oddari
Andhra Pradesh	Dravidian	Vadde	Korava, Vadde
Karnataka	Dravidian	Korava	Korava, Oddari, Vadde, Yerukali
Andhra Pradesh	Dravidian	Yerukali	Korava, Yerukali
Tamil Nadu	Dravidian	Gugavellalar	Gugavellalar, Madiga
Andhra Pradesh	Dravidian	Yanidi	Hindumalayali, Yanidi
Tamil Nadu	Dravidian	Hindumalayali	Hindumalayali, Yanidi
Andhra Pradesh	Dravidian	Madiga	Gugavellalar
Andhra Pradesh	Dravidian	Budagajangam	Budagajangam, Vysya
Andhra Pradesh	Dravidian	Vysya	Budagajangam, Vysya
Andhra Pradesh	Dravidian	Bestha	Bestha, Pattapu_Kapu
Andhra Pradesh	Dravidian	Pattapu_Kapu	Bestha, Pattapu_Kapu
Andhra Pradesh	Dravidian	Irula	Irula, Malayan
Kerala	Dravidian	Malayan	Irula
Karnataka	Dravidian	Brahmin_Karnataka	Brahmin_Karnataka, Havik
Karnataka	Dravidian	Havik	Brahmin_Karnataka, Havik
Orissa	AustroAsiatic	Mohali	Bhumij_Jharkhand, Gorait, Lohra, Manjhi_Jharkhand, Mohali, Mohli, Panika_MP, Panika_Jharkhand
Orissa	AustroAsiatic	Bhumij_Orissa	Bhumij_Jharkhand, Bhumij_Orissa, Hojo, Ho_Orissa, Lohra, Manjhi_Jharkhand, Munda, Santhal
Orissa	AustroAsiatic	Ho_Orissa	Bhumij_Jharkhand, Bhumij_Orissa, Birhor, Hojo, Ho_Orissa, Kharia, Lohra, Munda, Santhal
Chhattisgarh	IndoEuropean	Khairwar	Asur, Birhor, Ho_Orissa, Khairwar, Manjhi_Jharkhand, Munda, Santhal
Jharkhand	AustroAsiatic	Hojo	Bhumij_Jharkhand, Bhumij_Orissa, Hojo, Ho_Orissa, Lohra, Manjhi_Jharkhand, Munda, Santhal
Jharkhand	AustroAsiatic	Asur	Asur, Bhumij_Orissa, Ho_Orissa, Khairwar
Jharkhand	AustroAsiatic	Bhumij_Jharkhand	Barela, Bhumij_Jharkhand, Bhumij_Orissa, Birhor, Hojo, Ho_Orissa, Mohali, Munda, Santhal
Jharkhand	AustroAsiatic	Santhal	Bhumij_Jharkhand, Bhumij_Orissa, Birhor, Hojo, Ho_Orissa, Khairwar, Lohra, Manjhi_Jharkhand, Munda, Santhal
Jharkhand	AustroAsiatic	Munda	Bhumij_Jharkhand, Bhumij_Orissa, Hojo, Ho_Orissa, Santhal
Jharkhand	AustroAsiatic	Birhor	Bhumij_Jharkhand, Birhor, Ho_Orissa, Khairwar, Manjhi_Jharkhand, Santhal
Jharkhand	AustroAsiatic	Gorait	Gorait, Lohra, Mohali, Mohli, Oraon
Jharkhand	IndoEuropean	Panika_MP	Chauhan, Mohali, Mohli, Panika_MP, Panika_Jharkhand
Madhya Pradesh	IndoEuropean	Panika_Jharkhand	Mohali, Panika_MP, Panika_Jharkhand
Madhya Pradesh	AustroAsiatic	Korku	Mawasi
Madhya Pradesh	Astroasiatic	Mawasi	Korku, Mawasi
Madhya Pradesh	IndoEuropean	Baiga	Baiga, Bharia
Madhya Pradesh	Dravidian	Bharia	Baiga, Bharia
Madhya Pradesh	IndoEuropean	Barela	Barela, Bhilala, Bhumij_Jharkhand, Bink
Madhya Pradesh	IndoEuropean	Bhilala	Barela, Bhilala
Chhattisgarh	IndoEuropean	Chauhan	Chauhan, Lohra, Mohli, Panika_MP
Jharkhand	IndoEuropean	Tanti	Lohra, Mohli
Jharkhand	IndoEuropean	Mohli	Chauhan, Gorait, Lohra, Manjhi_Jharkhand, Mohali, Mohli, Oraon, Panika_MP, Tanti
Jharkhand	IndoEuropean	Lohra	Chauhan, Gorait, Manjhi_Jharkhand, Mohli, Santhal, Tanti
Jharkhand	IndoEuropean	Manjhi_Jharkhand	Bhumij_Orissa, Birhor, Hojo, Lohra, Manjhi_Jharkhand, Mohali, Mohli, Santhal
Jharkhand	IndoEuropean	Oraon	Gorait, Mohli, Oraon

Supplementary Table 4. IBD sharing across groups. Groups with more than one match for high shared IBD across groups (greater than 3 times the IBD score of CEU and ~1/3 the founder event strength of Ashkenazi Jews).

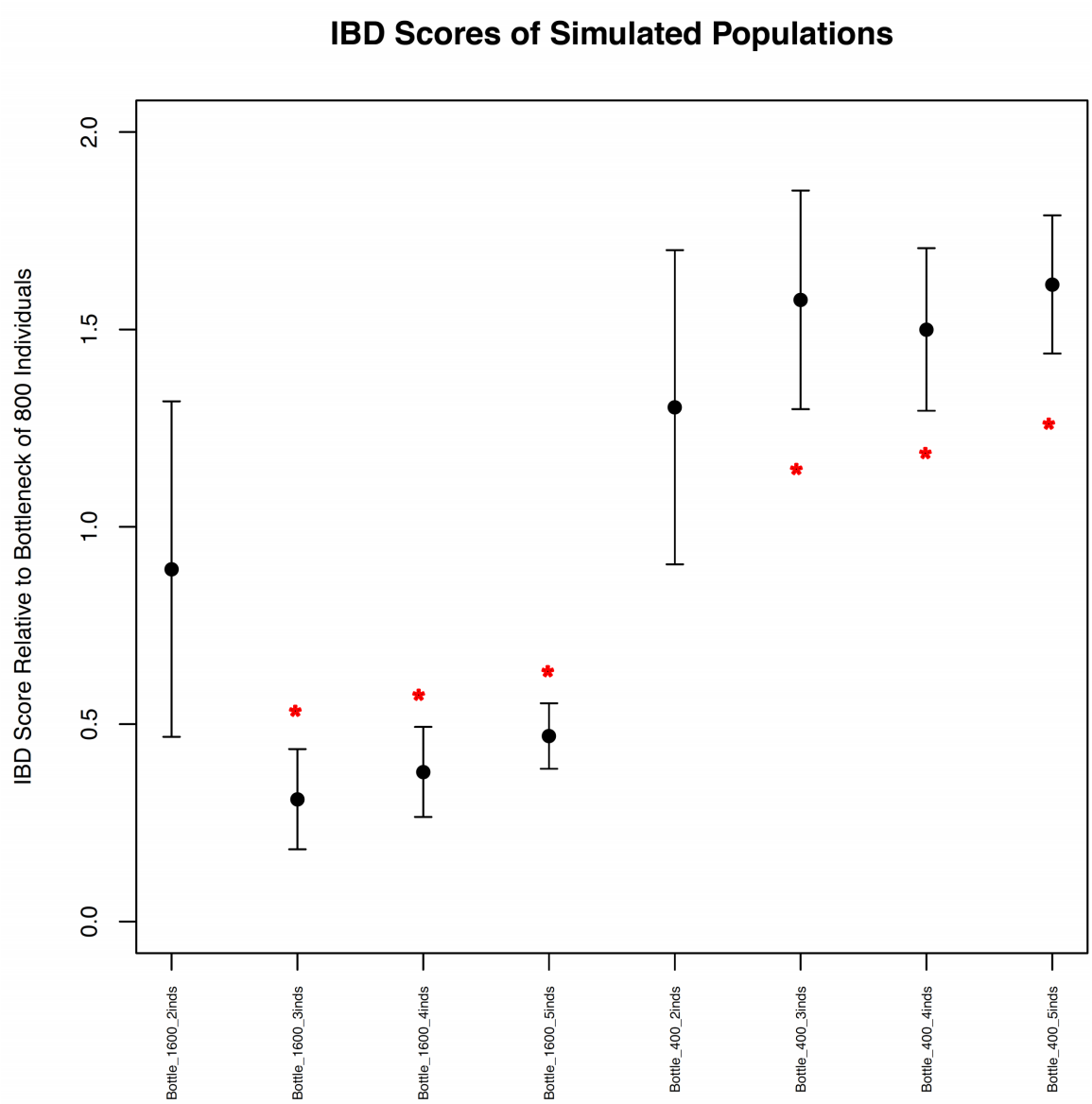
Source1	Source2	Target	f_3	Std. Err.	Z	SNPs	# Samples in Target
Mala	CHB	Tharu_Uttrakhand	-0.0135	0.0006	-21.1	92226	4
Mala	CHB	Tharu_UP	-0.0122	0.0007	-16.398	92159	3
Mala	CHB	Magar	-0.0097	0.0003	-36.818	81501	27
Mala	CHB	Newar	-0.0095	0.0005	-20.044	91534	6
Mala	CHB	Wan	-0.0089	0.0009	-10.359	92824	3
Mala	CHB	Syon	-0.0086	0.0008	-10.645	92535	3
Mala	CHB	Thakur	-0.0079	0.0004	-21.64	88899	10
Mala	CHB	Rajbanshi	-0.0067	0.0003	-20.85	89025	17
Mala	CHB	Sherpa	-0.0057	0.0009	-6.425	93939	3
Mala	CHB	Hazara	-0.0057	0.0004	-16.087	87624	14
Mala	CHB	Shah	-0.0049	0.0007	-7.021	92042	4
Mala	CHB	BEB (Bengali)	-0.0046	0.0002	-21.636	79064	86
Mala	CHB	Bink	-0.0040	0.0008	-4.906	91702	3
Mala	CHB	Chamar_UP	-0.0027	0.0013	-2.152	92603	2
Mala	CHB	Scheduled_Caste_Uttrakhand	-0.0024	0.0007	-3.542	92008	4
Mala	CHB	Minero	0.0028	0.0004	7.684	87579	18
Mala	CHB	Brahmin_Uttrakhand	0.0045	0.0006	7.462	89908	6
Mala	CHB	Burusho	0.0107	0.0004	25.374	84324	23
Mala	CHB	Kusunda	0.0127	0.0006	21.111	92036	10
Mala	CHB	Nyshi	0.0130	0.0008	16.768	91472	5
Mala	CHB	Nagaseema	0.0130	0.0011	11.796	91060	3

Supplementary Table 6. Groups with significant evidence of East Asian related admixture. Table of statistics of the form f_3(Target; Mala, CHB). Negative statistics indicate that the target group descends from an admixture of Mala (proxy for ANI and ASI ancestry), and CHB (proxy for East Asian-like ancestry). Standard errors are based on a weighted Block Jackknife (see Online Methods).



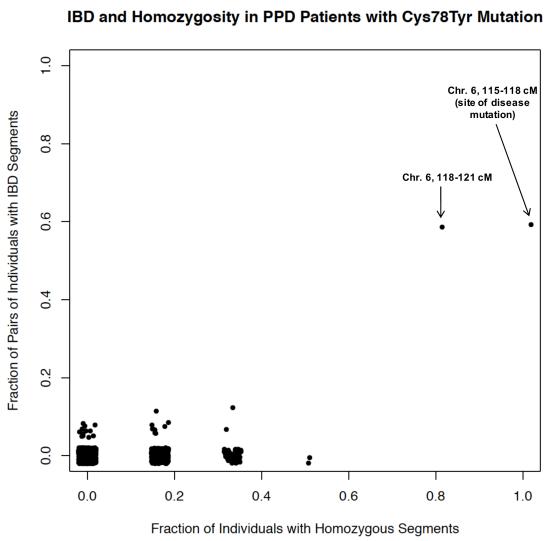
Comparison Group	Correlation (r^2)
Affymetrix 6.0 and Illumina	0.976
Affymetrix 6.0 and Human Origins	0.941
Affymetrix 6.0 and Illumina_Omni	0.982
Human Origins and Illumina	0.937
Human Origins and Illumina_Omni	0.962
Illumina and Illumina_Omni	0.987

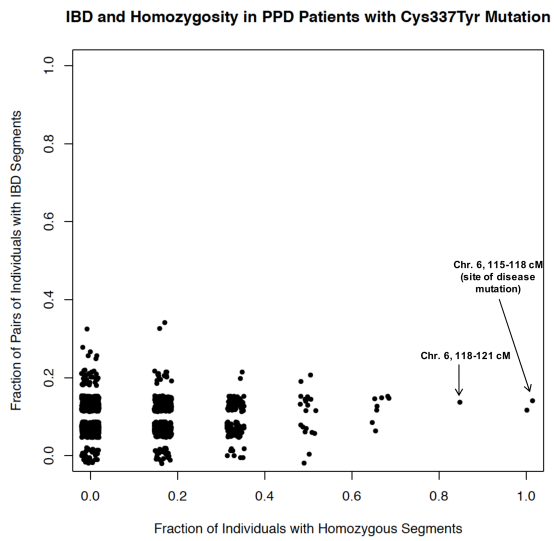
Supplementary Figure 1. Platform differences in IBD score. IBD Scores for 1000 Genomes groups in all 4 genotyping platforms after normalizing to FIN in each dataset.

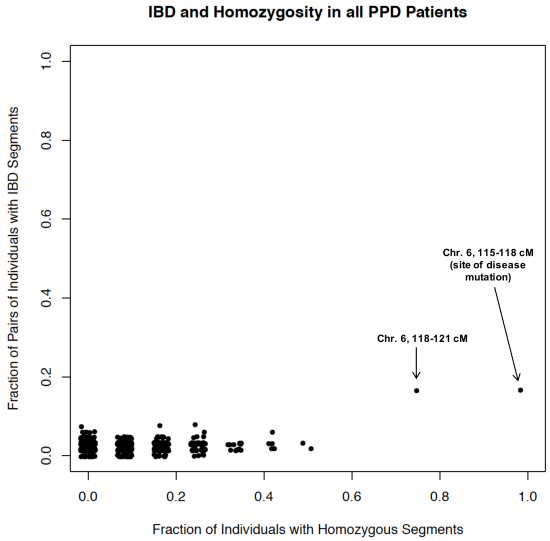


Supplementary Figure 2. Example simulation to determine number of samples needed to detect accurately a strong founder event. Scatterplot of IBD scores of simulated groups with bottlenecks twice as strong (400 individuals) or half as strong (1600 individuals) as a group with a bottleneck of strength similar to that of Finns (800 individuals). The points are the IBD scores of the different groups relative to those of the group with Finnish bottleneck strength. Error bars are standard errors. Red stars indicate significance defined as having a 95% confidence interval (±1.96*s.e.) that does not overlap with 1. Supplementary Table 2 shows the results for all 100 simulations.

123 A) B) C)

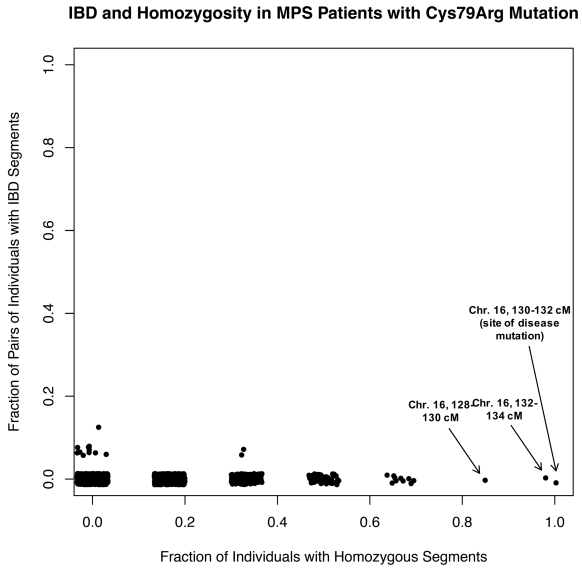






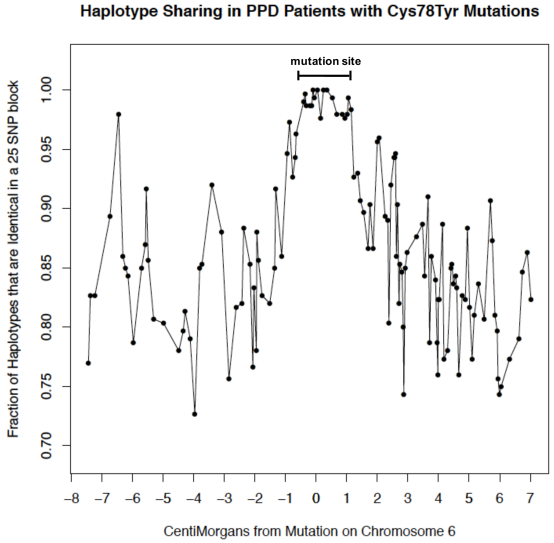
124
125

D)

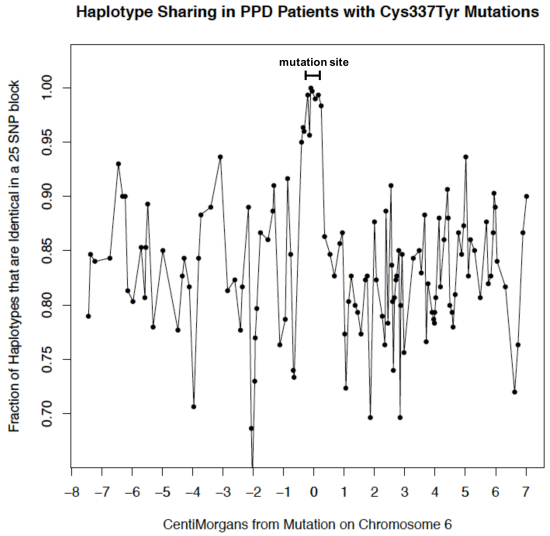


Supplementary Figure 3. Genomic differences between different progressive pseudorheumatoid dysplasia (PPD) and mucopolysaccharidosis (MPS) type IVA mutations. Scatterplots of percentage of individuals with homozygous segments vs. percentage of pairs of individuals with IBD segments (using our conservative thresholds for declaring IBD) for all 3 cM regions in the genome in (A) patients with Cys78Tyr mutations, (B) patients with Cys337Tyr mutations, (C) all PPD patients combined (those with Cys78Tyr and Cys337Tyr mutations), and (D) MPS patients with Cys79Arg mutations (for these patients we used 2 cM segments). Random jitter was added to the points for clarity. Segments that fell into 2 or more regions were included in both.

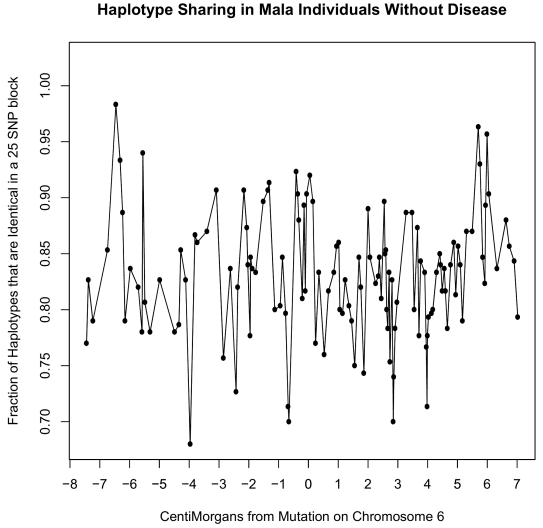
A)


Haplotype Sharing in PPD Patients with Cys78Tyr Mutations
mutation site
Fraction of Haplotypes that are Identical in a 25 SNP block
0.70
0.75
0.80
0.85
0.90
0.95
1.00
−8 −7 −6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6 7
CentiMorgans from Mutation on Chromosome 6

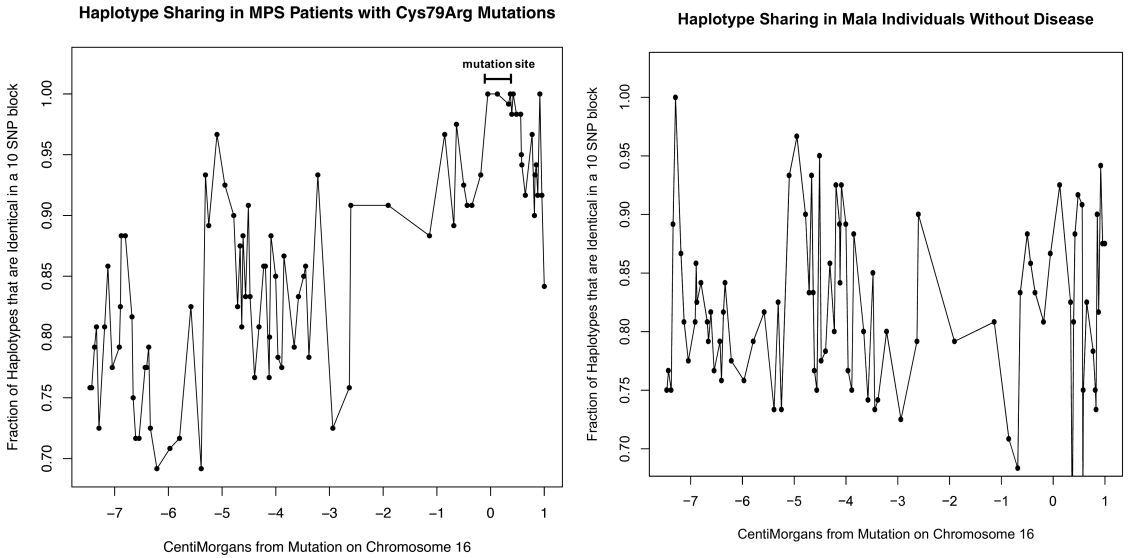
B)


Haplotype Sharing in PPD Patients with Cys337Tyr Mutations
mutation site
Fraction of Haplotypes that are Identical in a 25 SNP block
0.70
0.75
0.80
0.85
0.90
0.95
1.00
−8 −7 −6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6 7
CentiMorgans from Mutation on Chromosome 6

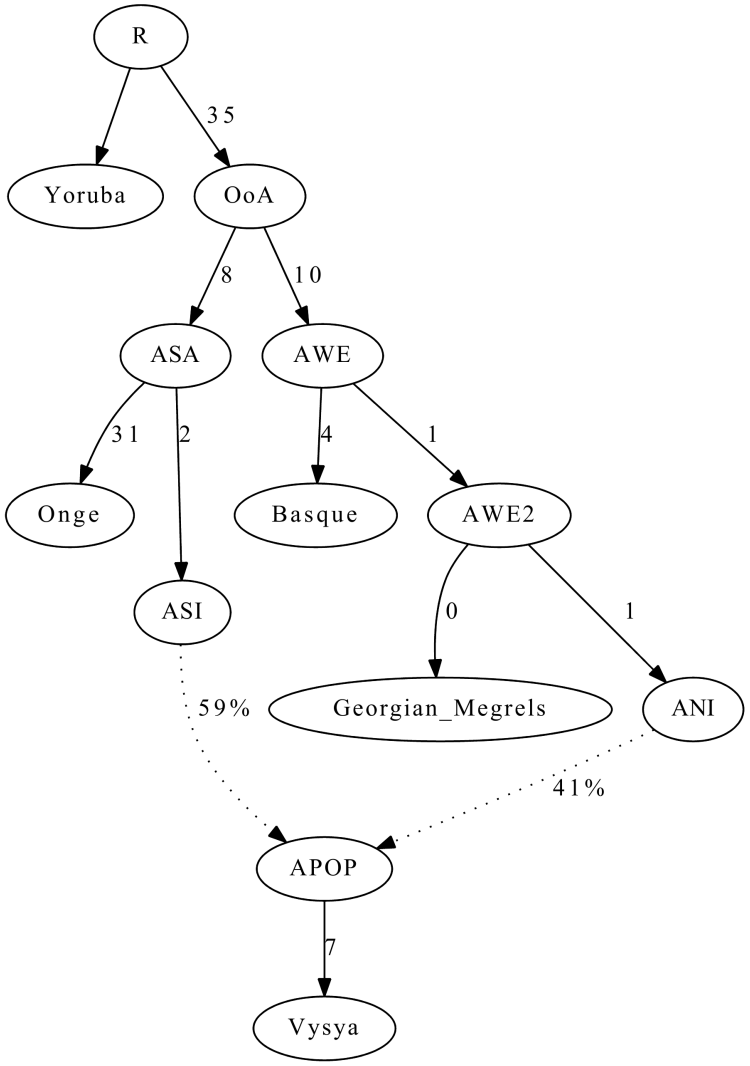
C)


Haplotype Sharing in Mala Individuals Without Disease
Fraction of Haplotypes that are Identical in a 25 SNP block
0.70
0.75
0.80
0.85
0.90
0.95
1.00
−8 −7 −6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6 7
CentiMorgans from Mutation on Chromosome 6

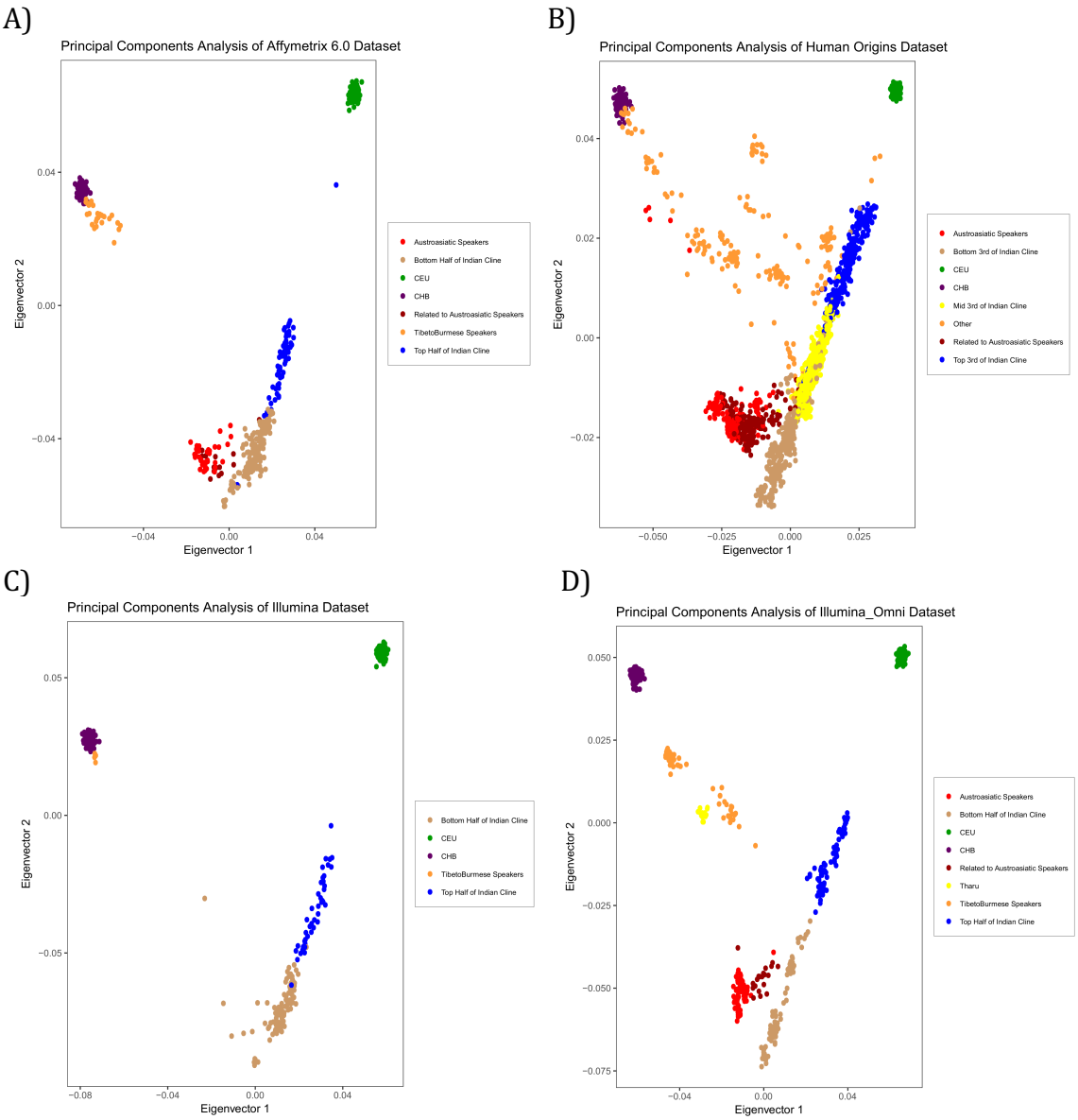
D) E)



Supplementary Figure 4. Haplotype sharing profiles of progressive pseudorheumatoid dysplasia (PPD) and mucopolysaccharidosis (MPS) type IVA patients. The fraction of haplotypes in 25 SNP blocks that are identical is plotted in the region surrounding the relevant mutation site on Chromosome 6 for MPS patients with A) Cys78Tyr mutations and B) Cys337Tyr mutations, as well as C) in the same region for Mala individuals without disease as a negative control. For MPS patients the mutation is near the end of the chromosome where SNP coverage is lower. Thus, the fraction of haplotypes in 10 SNP blocks that are identical is plotted in the region surrounding the relevant mutation site on Chromosome 16 for MPS patients with D) Cys79Arg mutations as well as E) in the same region for Mala individuals without disease.

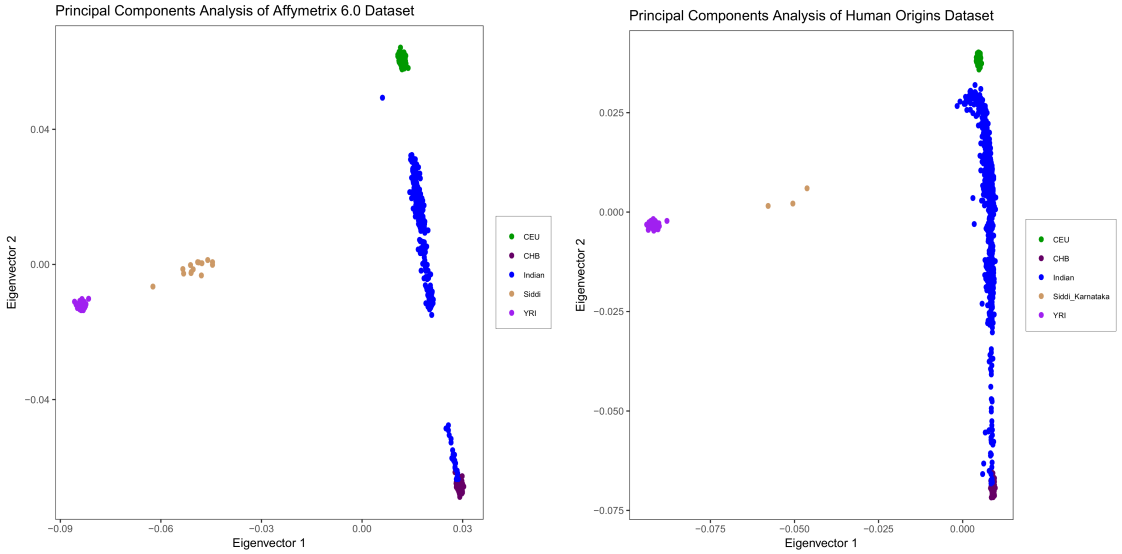


Supplementary Figure 5. Model used for estimating group specific drift in Indian groups (Vysya is an example Indian group). R=root; OoA=out of Africa; ASA=ancestral South Asian; ASI=ancestral Southern Indian; AWE=ancestral West Eurasian; ANI=ancestral North Indian; APOP=ancestral Indian group. Branch lengths are shown in units of F_{ST} x 1,000.



Supplementary Figure 6. Principal Components Analyses subdivided into clusters for F_{ST} analyses. (A) Affymetrix 6.0, (B) Human Origins, (C) Illumina, and (D) Illumina_Omni datasets. These plots are used to separate the groups into different clusters for F_{ST} analyses with the different sections of the Indian cline (e.g. "Top Half of Indian Cline" or "Bottom 3rd of Indian Cline") or Austroasiatic related (combined) each representing one cluster. (These analyses are described more fully in Online Methods section). 'Other' refers to ancestry outliers and groups with East Asian affinities.

A) B)



Supplementary Figure 7. Principal Components Analysis of A) Affymetrix 6.0 and B) Human Origins data along with CEU (European Americans), CHB (Han Chinese), and YRI (West Africans).