Supplementary Note

Inherited *GATA3* variants are associated with Ph-like childhood acute lymphoblastic leukemia and risk of relapse

[1]Virginia Perez-Andreu , [2]Kathryn G. Roberts, [3]Richard C. Harvey, [1]Wenjian Yang, [4]Cheng Cheng, [4]Deqing Pei, [1]Heng Xu, [5,6]Julie Gastier-Foster, [1]Shuyu E, [1,7]Joshua Yew-Suang Lim, [3]I-Ming Chen, [8]Yiping Fan, [9]Meenakshi Devidas, [10]Michael J. Borowitz, [1]Colton Smith, [11]Geoffrey Neale, [12]Esteban G. Burchard, [12]Dara G. Torgerson, [13]Federico Antillon Klussmann, [13]Cesar Rolando Najera Villagran, [14]Naomi J. Winick, [15]Bruce M. Camitta, [16]Elizabeth Raetz, [17]Brent WoodM.D., [18]Feng Yue, [16]William L. Carroll, [19]Eric Larsen, [20]W. Paul Bowman, [21]Mignon L. Loh, [22]Michael Dean, [23]Deepa Bhojwani, [23]Ching-Hon Pui, [1]William E. Evans, [1]Mary V. Relling, [24]Stephen P. Hunger, [3]Cheryl L. Willman, [2]Charles G. Mullighan, [1]Jun J. Yang

[1]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, TN, USA; [2]Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN, USA; [3]Cancer Center, University of New Mexico, Albuquerque, NM, USA;[4]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA; [5]Department of Pathology and Laboratory Medicine, Nationwide Children's Hospital, Columbus, OH, USA; [6]Departments of Pediatrics, Ohio State University School of Medicine, Columbus, OH, USA; [7]Department of Pediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore; [8]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, USA; [9]Department of Epidemiology and Health Policy Research, University of Florida, Gainesville, FL, USA;[10]Johns Hopkins Medical Institute, Baltimore, MD, USA;[11]Hartwell Center for Bioinformatics & Biotechnology, St. Jude Children's Research Hospital, Memphis, TN, USA; [12]Department of Bioengineering & Therapeutic Science and Medicine, University of California at San Francisco, San Francisco, CA, USA;[13]Unidad Nacional de Oncologia Pediatrica, Guatemala City, Guatemala; [14]Department of Pediatric Hematology/Oncology, University of Texas Southwestern Medical Center, Dallas, TX, USA; [15]Department of Pediatrics, Medical College of Wisconsin, Milwaukee, WI, USA; [16]New York University Cancer Institute, New York, NY, USA; [17]Department of Laboratory Medicine, University of Washington, Seattle, WA, USA; [18]Ludwig Institute for Cancer Research, University of California at San Diego School of Medicine, La Jolla, CA, USA;[19]Maine Children's Cancer Program, Scarborough, ME, USA,[20]Cook Children's Medical Center, Ft Worth, TX, USA;[21]Department of Pediatrics, University of California at San Francisco, San Francisco, CA, USA;[22]Laboratory of Experimental Immunology, National Cancer Institute, Frederick. MD, USA;[23]Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN, USA;[24] Children's Hospital Colorado, University of Colorado, Aurora, CO, USA.

Corresponding author

Jun J. Yang PhD
Dept. of Pharmaceutical Sciences, MS 313
St. Jude Children's Research Hospital
262 Danny Thomas Place, Memphis, TN 38105-3678
Email address: jun.yang@stjude.org
Phone: (901) 595-2517, FAX: (901) 595-8869

Acknowledgements for dbGaP datasets

MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159 through N01-HC-95169 and RR-024156. MESA, MESA Family, and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159 through N01-HC-95169 and RR-024156.

Funding support for the Genome-Wide Association of Schizophrenia Study was provided by the National Institute of Mental Health (R01 MH67257, R01 MH59588, R01 MH59571, R01 MH59565, R01 MH59587, R01 MH60870, R01 MH59566, R01 MH59586, R01 MH61675, R01 MH60879, R01 MH81800, U01 MH46276, U01 MH46289 U01 MH46318, U01 MH79469, and U01 MH79470) and the genotyping of samples was provided through the Genetic Association Information Network (GAIN). The datasets used for the analyses described in this manuscript were obtained from the database of Genotype and Phenotype database (dbGaP) found at http://www.ncbi.nlm.nih.gov/gap through dbGaP (accession number phs000021.v1.p1). Samples and associated phenotype data for the Genome-Wide Association of Schizophrenia Study were provided by the Molecular Genetics of Schizophrenia Collaboration (Principal Investigator: P.V. Gejman, Evanston Northwestern Healthcare (ENH) and Northwestern University, Evanston, Illinois).

Funding support for the Whole Genome Association Study of Bipolar Disorder was provided by the US National Institute of Mental Health (NIMH) and the genotyping of samples was provided through the Genetic Association Information Network (GAIN). The datasets used for the analyses described in this manuscript were obtained from dbGaP (accession number phs000017.v1.p1). Samples and associated phenotype data for the Collaborative Genomic Study of Bipolar Disorder were provided by The NIMH Genetics Initiative for Bipolar Disorder. Data and biomaterials were collected in four projects that participated in the NIMH Bipolar Disorder Genetics Initiative. From 1991 to 1998, the principal investigators and co-investigators were: Indiana University, Indianapolis,

Indiana, U01 MH46282, J. Nurnberger, M. Miller and E. Bowman; Washington University, St. Louis, Missouri, U01 MH46280, T. Reich, A. Goate and J. Rice; Johns Hopkins University, Baltimore, Maryland, U01 MH46274, J.R. DePaulo Jr., S. Simpson and C. Stine; NIMH Intramural Research Program, Clinical Neurogenetics Branch, Bethesda, Maryland, E. Gershon, D. Kazuba and E. Maxwell. Data and biomaterials were collected as part of ten projects that participated in the NIMH Bipolar Disorder Genetics Initiative. From 1999 to 2003, the principal investigators and co-investigators were: Indiana University, Indianapolis, Indiana, R01 MH59545, J. Nurnberger, M.J. Miller, E.S. Bowman, N.L. Rau, P.R. Moe, N. Samavedy, R. El-Mallakh (University of Louisville, Louisville, Kentucky), H. Manji (Wayne State University, Detroit, Michigan), D.A. Glitz (Wayne State University), E.T. Meyer, C. Smiley, T. Foroud, L. Flury, D.M. Dick and H. Edenberg; Washington University, St. Louis, Missouri, R01 MH059534, J. Rice, T. Reich, A. Goate and L. Bierut; Johns Hopkins University, Baltimore, Maryland, R01 MH59533, M. McInnis, J.R. DePaulo Jr., D.F. MacKinnon, F.M. Mondimore, J.B. Potash, P.P. Zandi, D. Avramopoulos and J. Payne; University of Pennsylvania, Philadelphia, Pennsylvania, R01 MH59553, W. Berrettini; University of California, Irvine, California, R01 MH60068, W. Byerley and M. Vawter; University of Iowa, Iowa City, Iowa, R01 MH059548, W. Coryell and R. Crowe; University of Chicago, Chicago, Illinois, R01 MH59535, E. Gershon, J. Badner, F. McMahon, C. Liu, A. Sanders, M. Caserta, S. Dinwiddie, T. Nguyen and D. Harakal; University of California, San Diego, California, R01 MH59567, J. Kelsoe and R. McKinney; Rush University, Chicago, Illinois, R01 MH059556, W. Scheftner, H.M. Kravitz, D. Marta, A. Vaughn-Brown and L. Bederow; NIMH Intramural Research Program, Bethesda, Maryland, 1Z01MH002810-01, F.J. McMahon, L. Kassem, S. Detera-Wadleigh, L. Austin and D.L. Murphy.

Supplementary Table 1. Multivariate analysis of rs3824662 and rs3781093 for association with Ph-like ALL in COG AALL0232.

						Ph-like ALL *vs.* non-Ph-like ALL				Ph-like ALL *vs.* non-ALL controls			
Gene	***rs ID***	***Allele A***	***Allele B***	***Position (hg19)***	***SNP type***	***P-value[1]***	***OR (95%, CI)***	***P-value[2]***	***OR (95%, CI)***	***P-value[3]***	***OR (95%, CI)***	***P-value[4]***	***OR (95%, CI)***
GATA3	**rs3781093**	T	C	8101927	Genotyped	2.62×10^{-7}	3.09 (2.06-4.63)	0.306684	1.53 (0.66-3.51)	4.94×10^{-12}	3.70 (2.61-5.25)	0.26976	1.58 (0.69-3.63)
GATA3	**rs3824662**	C	A	8104208	Genotyped	1.05×10^{-8}	3.17 (2.12-4.74)	0.008	3.09 (1.32-7.25)	2.17×10^{-14}	3.75 (2.65-5.30)	0.001	3.49 (1.61-7.55)

Abbreviations: OR, Odds ratio; CI, confidence interval.
Association of SNP genotype and Ph-like ALL was evaluated by logistic regression, after adjusting for genetic ancestry. Chromosomal locations are based on hg19.
P-value[1]: Ph-like ALL *vs.* non-Ph-like ALL adjusting for genetic ancestry
P-value[2] : Ph-like ALL *vs.* non-Ph-like ALL adjusting for genetic ancestry and *GATA3* SNPs genotype.
P-value[3] : Ph-like ALL *vs*. non-ALL controls adjusting for genetic ancestry
P-value[4] : Ph-like ALL *vs.* non-ALL controls adjusting for genetic ancestry and *GATA3* SNPs genotype.

Supplementary Table 2. Clinical characteristics of patients included in this study by cohort.

	Children's Oncology Group Cohort		
Treatment Protocol	**AALL0232**	**P9906**	**P9905**
Number of patients	**(n=511)**	**(n=215)**	**(n=889)**
Race[a]			
Asian	17 (3.0)	3 (1.3)	19 (2.1)
African-American	25 (5.0)	14 (6.5)	51 (5.7)
Hispanic	149 (29.1)	53 (24.6)	191 (21.5)
European-American	160 (31.3)	116 (54.0)	520 (58.5)
Sex			
Female	237 (46.4)	68 (31.6)	452 (50.8)
Male	273 (53.5)	147 (68.3)	437 (49.2)
Missing	1 (0.1)	NA	NA
Age at diagnosis, y			
<10	222 (43.5)	73 (34.0)	690 (77.6)
≥10	288 (56.4)	142 (66.0)	198 (22.3)
Missing	1 (0.1)	NA	1 (0.11)
Leucocyte count at diagnosis, /µL			
<50 000	223 (43.7)	120 (56.0)	737 (82.9)
≥50 000	287 (56.2)	95 (44.1)	152 (17.1)
Missing	1 (0.1)	NA	NA
CNS status			
CNS3 or traumatic	11 (2.5)	28 (13.0)	0 (0.0)
CNS 1	397 (81.1)	163 (76.0)	795 (89.4)
CNS 2	81 (16.0)	24 (11.1)	93 (10.5)
Missing	22 (4.3)	NA	1 (0.11)

Data is presented as No. (%) unless otherwise indicated.
[a]Genetic ancestry was determined by using STRUCTURE. Asians, African-Americans, Hispanics, and European-Americans were identified as >90%, >70%, >10% and higher than African ancestry, >95% of Asian, African, Native-American and European genetic ancestry, respectively.
NA: Not applicable.

Supplementary Table 3A. List of primers for Sanger sequencing of *GATA3* SNPs (rs3824662 and rs3781093).

***GATA 3* SNPs**	**PCR Primers**		**Sanger sequencing Primers**
	Forward	**Reverse**	
rs3824662	5'-TATCACCCTCCCCACCA	5'-GGAAAGCCCCAGATCAA	5'-TATCACCCTCCCCACCA
rs3781093	5'-TTCCTGTGCTCTGTTCCTT	5'-GGCTCAGGATAAACAATG	5'-TTCCTGTGCTCTGTTCCTT

Supplementary Table 3B. List of primers for real-time PCR of *GATA3* mRNA

***GATA3* Forward Primer**	5'-TCACAAAATGAACGGACAGAACC-3'
***GATA3* Reverse Primer**	5'-CAGCCTTCGCTTGGGCTTAAT-3'

Supplementary Table 4. Association of germline *JAK2* SNPs with somatic *JAK2* mutation

rsID	*P-value*[1]	OR (95%, CI)	*P-value*[2]	OR (95%, CI)
rs2149556	0.1978	1.35 (0.84-2.15)	0.1166	1.41 (0.91-2.19)
rs7864782	0.2329	1.30 (0.83-2.03)	0.1703	1.34 (0.87-2.05)
rs10815144	0.2584	0.76 (0.48-1.21)	0.2044	0.75 (0.48-1.16)
rs10124001	0.2854	1.97 (0.56-6.94)	0.3740	1.70 (0.51-5.59)
rs10119004	0.4588	0.84 (0.53-1.32)	0.3488	0.81 (0.52-1.25)
rs10974944	0.6754	0.89 (0.54-1.48)	0.6591	0.89 (0.54-1.47)
rs11793659	0.7882	1.06 (0.66-1.71)	0.7556	1.07 (0.67-1.72)
rs1327493	0.8624	1.10 (0.35-3.42)	0.9095	1.06 (0.37-3.02)
rs17425637	0.9616	1.01 (0.62-1.64)	0.9505	1.01 (0.63-1.63)
rs12340895	0.9661	0.98 (0.59-1.63)	0.9390	0.98 (0.59-1.60)
rs6476934	0.9921	1.00 (0.23-4.40)	0.7303	1.28 (0.30-5.30)

Association between *JAK2* somatic lesion and J*AK2* germline SNPs was tested in the combined cohort (COG AALL0232, COG P9906 and COG P9905) and in the discovery non-ALL control group (N=6,661), by logistic regression after adjusting for genetic ancestry.
P-value[1] : comparing allele frequency between ALL with *vs.* without *JAK2* mutations
P-value[2] : comparing allele frequency between ALL with *JAK2* mutations *vs.* non-ALL controls
Abbreviations: OR, Odds Ratio; CI, confidence interval

Supplementary Table 5. Gene Set Enrichment Analysis of ALL cells overexpressing *GATA3 vs.* control

Nalm6				UOCB1			
Gene symbol	Rank Metric Score	Running ES	Core Enrichment	Gene symbol	Rank Metric Score	Running ES	Core Enrichment
ANXA1*	0.2595	0.3822	Yes	***ANXA1****	3.9373	0.1921	Yes
BCL6	0.6874	0.1169	Yes	***CASP10***	0.2628	0.4673	Yes
CASP10	0.2619	0.3630	Yes	***GBP2****	0.5453	0.3644	Yes
CD99*	0.4257	0.2728	Yes	***IL2RA****	0.2100	0.5066	Yes
CTHRC1*	0.1918	0.4009	Yes	***LCP2***	0.4846	0.4372	Yes
DOK4*	0.2165	0.3922	Yes	***LYZ***	0.5054	0.4138	Yes
ECM1*	0.1871	0.4130	Yes	***MYBL1****	0.2035	0.5128	Yes
GBP2*	0.6666	0.1664	Yes	***ANKRD28***	0.3527	0.4689	Yes
IL2RA*	0.1696	0.4312	Yes	***ANTXR2****	1.6885	0.2743	Yes
LCP2	0.3432	0.3258	Yes	***CD300A****	0.4056	0.4548	Yes
LYZ	0.3145	0.3479	Yes	***CDC42EP3****	0.5181	0.3893	Yes
MYBL1*	0.4396	0.2415	Yes	***ENAM****	0.2026	0.5223	Yes
NRXN3*	0.6131	0.2114	Yes	***MMRN1****	0.2250	0.4844	Yes
RAPGEF3*	0.4032	0.3023	Yes	***PON2***	0.2216	0.4934	Yes
SELL*	0.8998	0.0662	Yes	***PSTPIP2****	1.3373	0.3394	Yes
TTN	0.1702	0.4187	Yes	***S100Z****	0.1939	0.5244	Yes
ABCA9	0.0429	0.3377	No	***SERPINA1***	0.2173	0.5017	Yes
ABL1	-0.0621	-0.0746	No	***STON2***	0.2425	0.4831	Yes
AHR	-0.0154	0.0598	No	***TBXAS1****	0.2468	0.4733	Yes
ANKRD28	0.0419	0.3347	No	***BCL6***	-0.0679	0.0227	No
ANTXR2	0.0136	0.2152	No	***CD99***	-0.0369	0.0885	No
ANXA4	0.0272	0.2833	No	***CTHRC1***	-0.0742	0.0131	No
ATP10A	0.0513	0.3482	No	***DOK4***	-0.1730	-0.0245	No
B3GNTL1	-0.0376	-0.0216	No	***ECM1***	-0.0115	0.1645	No
BAALC	0.1278	0.4113	No	***NRXN3***	0.0116	0.2491	No
BSPRY	-0.0862	-0.1009	No	***RAPGEF3***	-0.0096	0.1706	No
BST1	-0.0242	0.0276	No	***SELL***	0.0753	0.4321	No
C1QTNF4	-0.0091	0.0879	No	***TTN***	0.0200	0.2757	No
CA6	0.0403	0.3357	No	***ABCA9***	-0.0732	0.0115	No
CASP1	-0.1761	-0.1232	No	***ABL1***	0.0543	0.3721	No
CAV1	0.0901	0.3956	No	***AHR***	-0.0090	0.1722	No
CCL17	0.0917	0.3925	No	***ANXA4***	0.0048	0.2253	No
CCND2	-0.0397	-0.0236	No	***ATP10A***	-0.0898	0.0054	No
CD300A	0.0414	0.3387	No	***B3GNTL1***	-0.1538	-0.0372	No
CD302	-0.0134	0.0677	No	***BAALC***	0.1046	0.4937	No
CDC42EP3	-0.0111	0.0788	No	***BSPRY***	-0.2403	-0.0231	No
CEACAM6	0.1038	0.4023	No	***BST1***	-0.0319	0.1004	No
CFD	0.0782	0.3889	No	***C1QTNF4***	0.0185	0.2704	No
CFP	0.0583	0.3540	No	***CA6***	0.1201	0.5037	No
CHN1	-0.3405	0.0038	No	***CASP1***	-0.1978	-0.0254	No
CHN2	0.0431	0.3355	No	***CAV1***	-0.0197	0.1393	No
CHRNA1	-0.2642	-0.0784	No	***CCL17***	-0.0075	0.1774	No
CRADD	0.0441	0.3308	No	***CCND2***	-0.0671	0.0213	No
CSTA	0.1191	0.4095	No	***CD302***	0.1744	0.5147	No
CTDSPL	-0.1187	-0.1181	No	***CEACAM6***	0.1266	0.5041	No
CYYR1	0.0535	0.3452	No	***CFD***	0.1061	0.4867	No
DFNA5	0.0440	0.3338	No	***CFP***	-0.0600	0.0324	No
DPYD	0.0729	0.3828	No	***CHN1***	-0.0538	0.0476	No
DUSP6	-0.0355	-0.0157	No	***CHN2***	-0.0228	0.1293	No
EGFL7	0.0858	0.3926	No	***CHRNA1***	0.0218	0.2801	No
EMP1	-0.0144	0.0640	No	***CRADD***	-0.1591	-0.0280	No
ENAM	0.0495	0.3431	No	***CSTA***	-0.0421	0.0771	No

Supplementary Table 5. Gene Set Enrichment Analysis of ALL cells overexpressing *GATA3 vs.* control

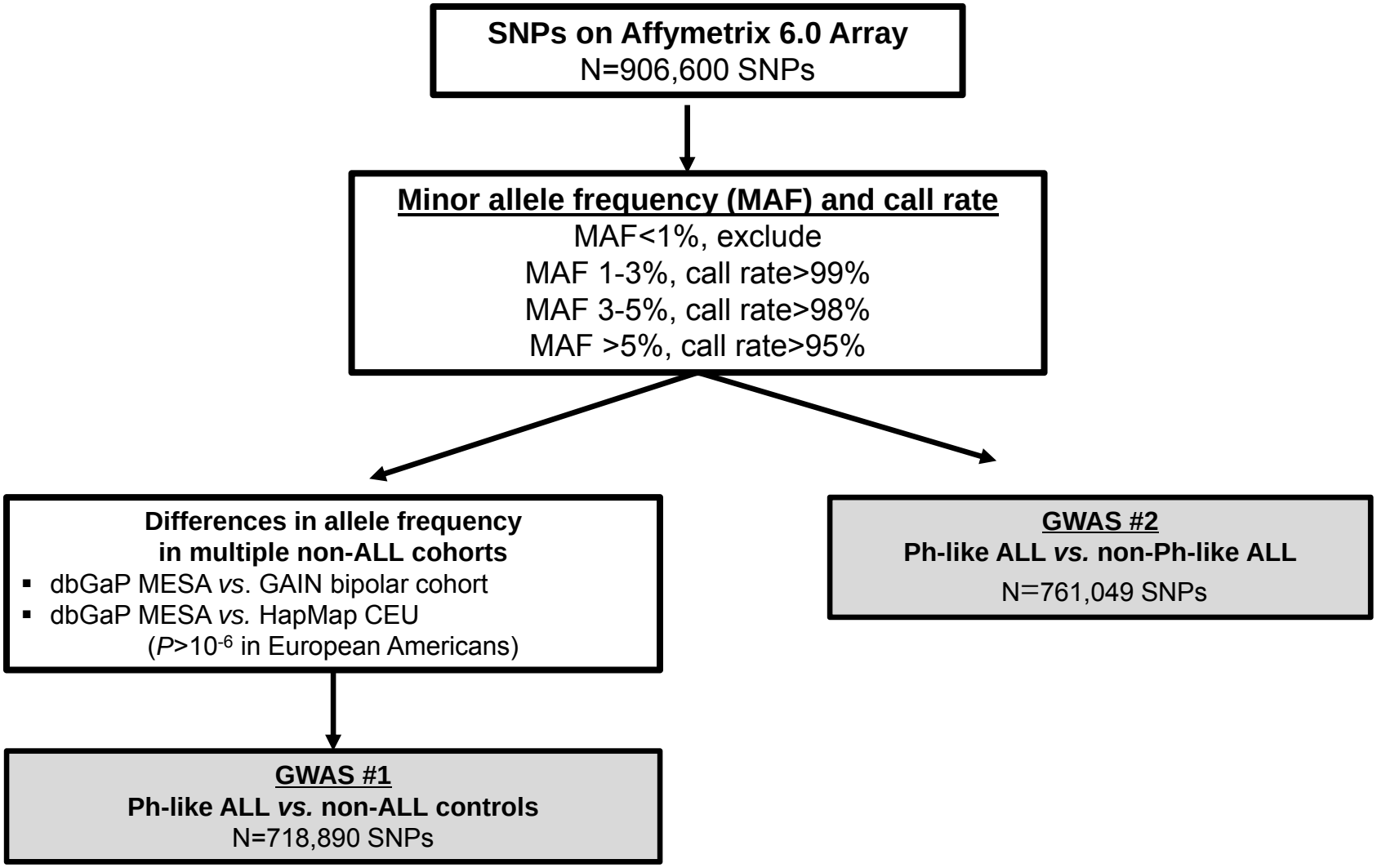
FARP1	0.0014	0.1468	No	***CTDSPL***	0.0202	0.2751	No
FBXW7	0.0151	0.2220	No	***CYYR1***	0.0955	0.4788	No
FSCN1	-0.0196	0.0448	No	***DFNA5***	0.1338	0.4976	No
FUT7	0.0688	0.3732	No	***DPYD***	0.1284	0.5012	No
GADD45A	-0.0849	-0.1047	No	***DUSP6***	-0.0181	0.1437	No
GBP5	-0.0004	0.1359	No	***EGFL7***	-0.0895	-0.0028	No
GIMAP1	0.0255	0.2758	No	***EMP1***	0.0000	0.2066	No
GIMAP6	0.0653	0.3710	No	***FARP1***	0.0020	0.2141	No
GLIPR1	-0.0380	-0.0201	No	***FBXW7***	-0.0333	0.0973	No
GPR110	0.0091	0.1925	No	***FSCN1***	-0.0065	0.1811	No
GPR56	-0.0186	0.0482	No	***FUT7***	-0.1438	-0.0410	No
GYPC	-0.2552	-0.0971	No	***GADD45A***	0.0112	0.2483	No
HES1	0.0324	0.3073	No	***GBP5***	0.1386	0.4992	No
ID1	-0.0214	0.0398	No	***GIMAP1***	0.0068	0.2330	No
IFITM1	0.0521	0.3474	No	***GIMAP6***	-0.0152	0.1520	No
IFITM3	-0.0207	0.0415	No	***GLIPR1***	0.0347	0.3219	No
IGFBP7	0.0565	0.3503	No	***GPR110***	0.1500	0.5087	No
IPO11	-0.0831	-0.1068	No	***GPR56***	-0.0421	0.0751	No
KAZALD1	0.0149	0.2219	No	***GYPC***	-0.3484	0.0131	No
KBTBD8	-0.0616	-0.0780	No	***HES1***	0.0354	0.3224	No
KCNE3	-0.0062	0.1034	No	***ID1***	0.0408	0.3350	No
KLF9	-0.0309	0.0016	No	***IFITM1***	-0.1198	-0.0350	No
LIMS1	0.0471	0.3390	No	***IFITM3***	0.0504	0.3609	No
LST1	0.0176	0.2340	No	***IGFBP7***	-0.2980	-0.0117	No
MAPKAPK3	-0.1170	-0.1339	No	***IPO11***	-0.0529	0.0474	No
MCTP1	0.0631	0.3675	No	***KAZALD1***	-0.2275	-0.0298	No
MDFIC	-0.0131	0.0689	No	***KBTBD8***	-0.1121	-0.0274	No
MINA	-0.2657	-0.0588	No	***KCNE3***	0.0494	0.3608	No
MMP28	0.0080	0.1856	No	***KLF9***	-0.1294	-0.0418	No
MMRN1	-0.0053	0.1083	No	***LIMS1***	0.1334	0.5031	No
MS4A4A	-0.0184	0.0481	No	***LST1***	-0.0744	0.0160	No
MSRB3	-0.0403	-0.0234	No	***MAPKAPK3***	-0.0757	0.0158	No
MUC4	0.0122	0.2088	No	***MCTP1***	0.0181	0.2703	No
NFE2L2	0.0156	0.2239	No	***MDFIC***	-0.0157	0.1510	No
NPDC1	0.0440	0.3370	No	***MINA***	0.0160	0.2639	No
NT5E	0.0310	0.3013	No	***MMP28***	-0.0127	0.1607	No
NUDT4	-0.0501	-0.0541	No	***MS4A4A***	-0.1490	-0.0391	No
OLFML2A	0.0851	0.3978	No	***MSRB3***	0.1058	0.4914	No
OR7A5	0.0071	0.1811	No	***MUC4***	0.0418	0.3359	No
PELI1	-0.0601	-0.0782	No	***NFE2L2***	-0.0133	0.1589	No
PHACTR1	-0.1256	-0.1077	No	***NPDC1***	0.0222	0.2800	No
PON2	0.0334	0.3103	No	***NT5E***	-0.1546	-0.0305	No
PRX	0.0097	0.1947	No	***NUDT4***	-0.3040	0.0019	No
PSTPIP2	0.0383	0.3286	No	***OLFML2A***	-0.0050	0.1866	No
PTPN14	0.1094	0.4045	No	***OR7A5***	0.0306	0.3101	No
RASSF8	0.0760	0.3877	No	***PELI1***	0.1256	0.5087	No
ROBO3	-0.0014	0.1303	No	***PHACTR1***	0.0404	0.3353	No
ROBO4	0.0079	0.1859	No	***PRX***	0.0332	0.3178	No
RRAS	-0.0565	-0.0714	No	***PTPN14***	0.0041	0.2225	No
S100A8	-0.1265	-0.0994	No	***RASSF8***	-0.0377	0.0878	No
S100Z	-0.0623	-0.0705	No	***ROBO3***	0.0867	0.4590	No
SCHIP1	0.0829	0.3969	No	***ROBO4***	0.0173	0.2682	No
SERPINA1	-0.0707	-0.0891	No	***RRAS***	-0.0357	0.0913	No
SH3BP5	-0.0021	0.1259	No	***S100A8***	-0.1936	-0.0319	No
SLC2A5	-0.2737	-0.0393	No	***SCHIP1***	-0.0822	0.0082	No

Supplementary Table 5. Gene Set Enrichment Analysis of ALL cells overexpressing *GATA3 vs.* control

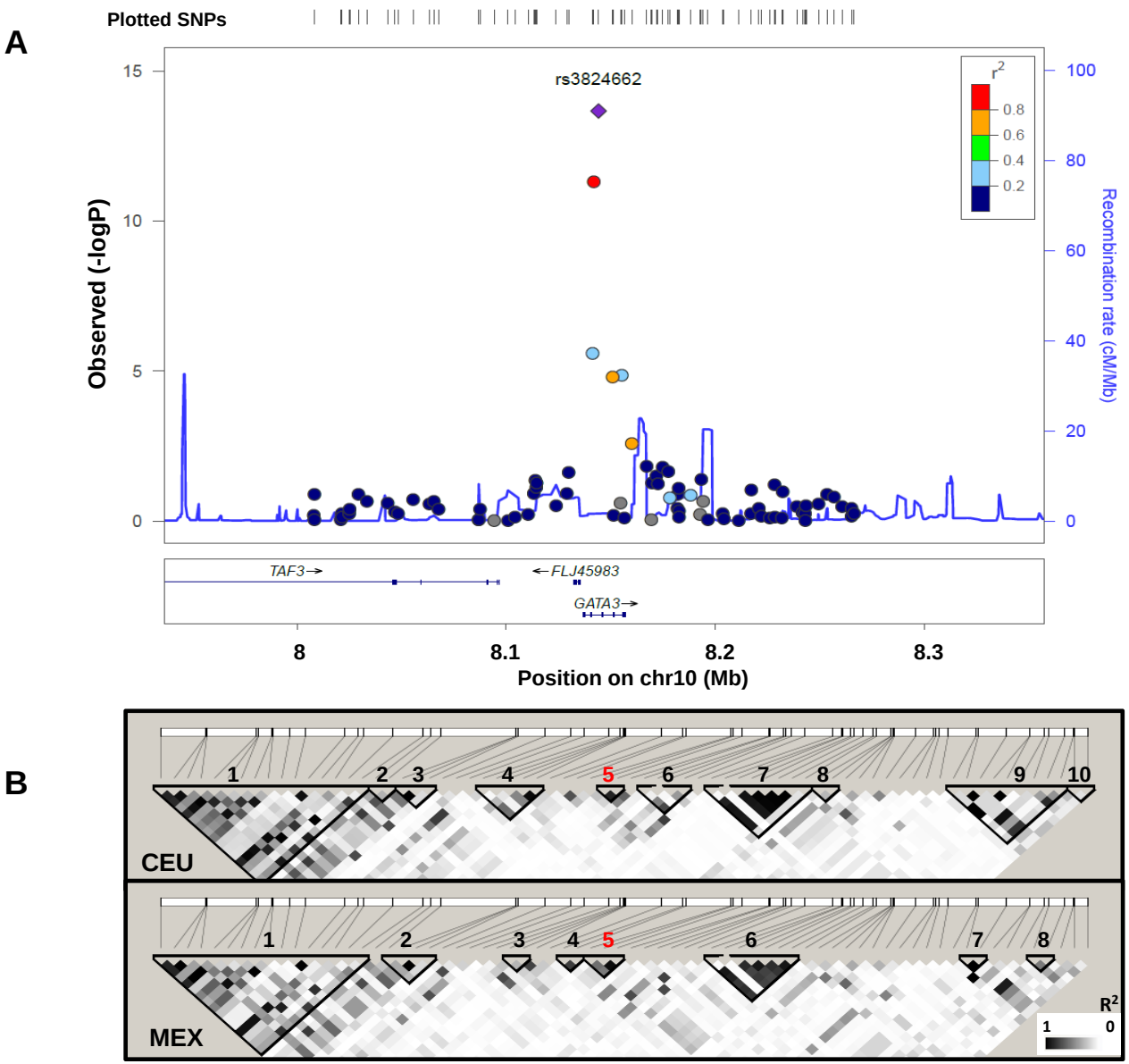
SLC39A10	-0.2173	-0.1092	No	***SH3BP5***	0.0360	0.3230	No
SLC44A1	0.0134	0.2154	No	***SLC2A5***	0.0072	0.2344	No
SOCS2	-0.0105	0.0811	No	***SLC39A10***	-0.0762	0.0191	No
SPARC	-0.1224	-0.1135	No	***SLC44A1***	-0.0042	0.1897	No
SPON1	0.0445	0.3295	No	***SOCS2***	-0.1358	-0.0379	No
STAB1	0.0545	0.3455	No	***SPARC***	-0.2845	-0.0232	No
STON2	-0.1032	-0.1246	No	***SPON1***	-0.1067	-0.0241	No
SUSD3	0.1036	0.4098	No	***STAB1***	0.0628	0.3966	No
SV2A	0.0249	0.2734	No	***SUSD3***	-0.1684	-0.0291	No
TBXAS1	0.0678	0.3752	No	***SV2A***	-0.0668	0.0189	No
THBS1	0.0369	0.3266	No	***THBS1***	0.0370	0.3249	No
TMEM154	-0.0787	-0.1026	No	***TMEM154***	0.1572	0.5112	No
TTYH2	-0.1903	-0.1156	No	***TTYH2***	-0.1322	-0.0389	No
UACA	-0.1186	-0.1267	No	***UACA***	0.1635	0.5115	No
UPP1	0.0371	0.3251	No	***UPP1***	-0.0896	0.0015	No

GSEA analysis comparing ALL cells ectopically overexpressing *GATA3* vs. cells transduced with empty vector, with PAM-based Ph-like signature (upregulated) as the *a priori* gene set.

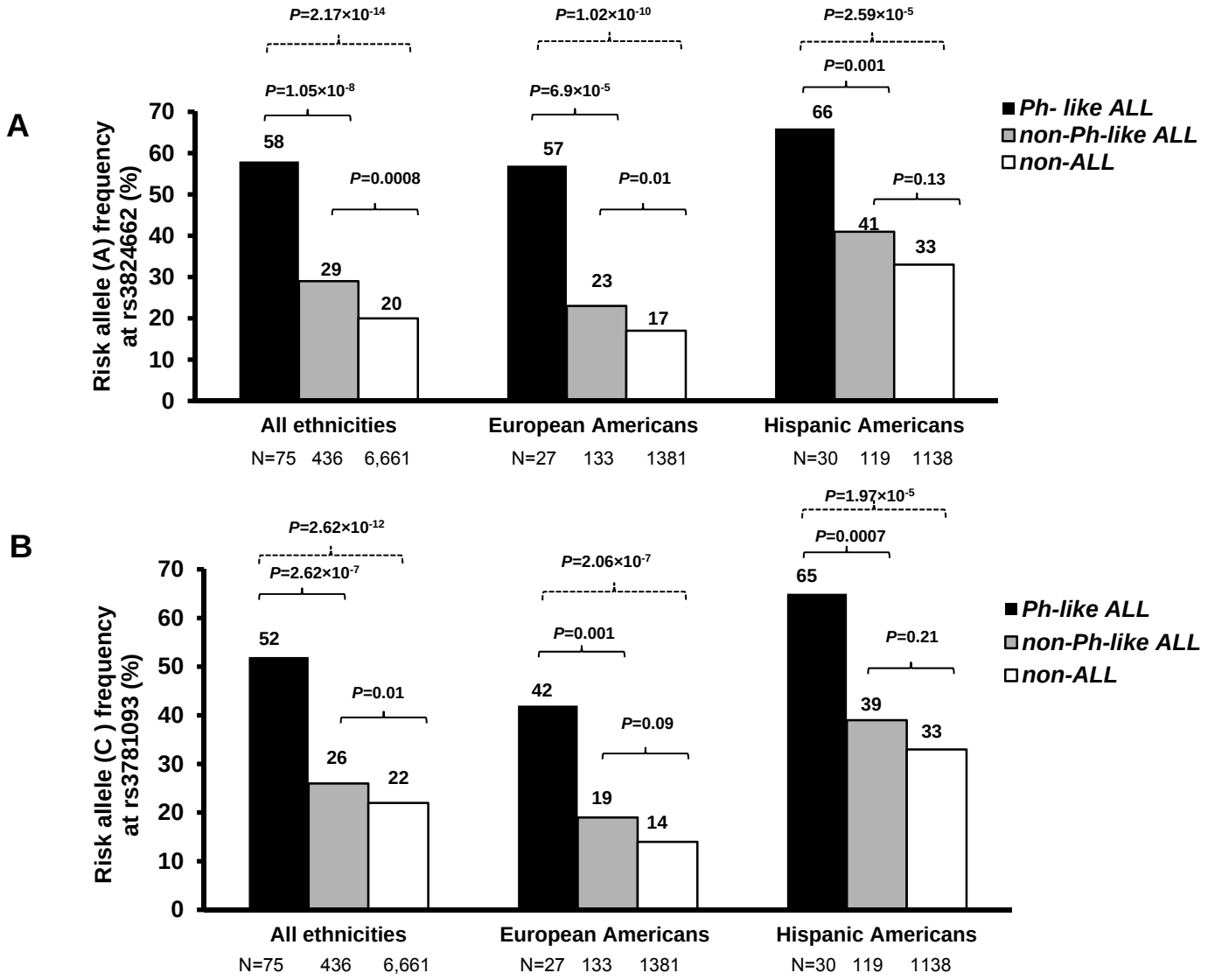
*Denotes genes that were upregulated in patients carrying the A allele at rs3824662 compared with those not carrying the A allele in both COG AALL0232 and COG P9906.



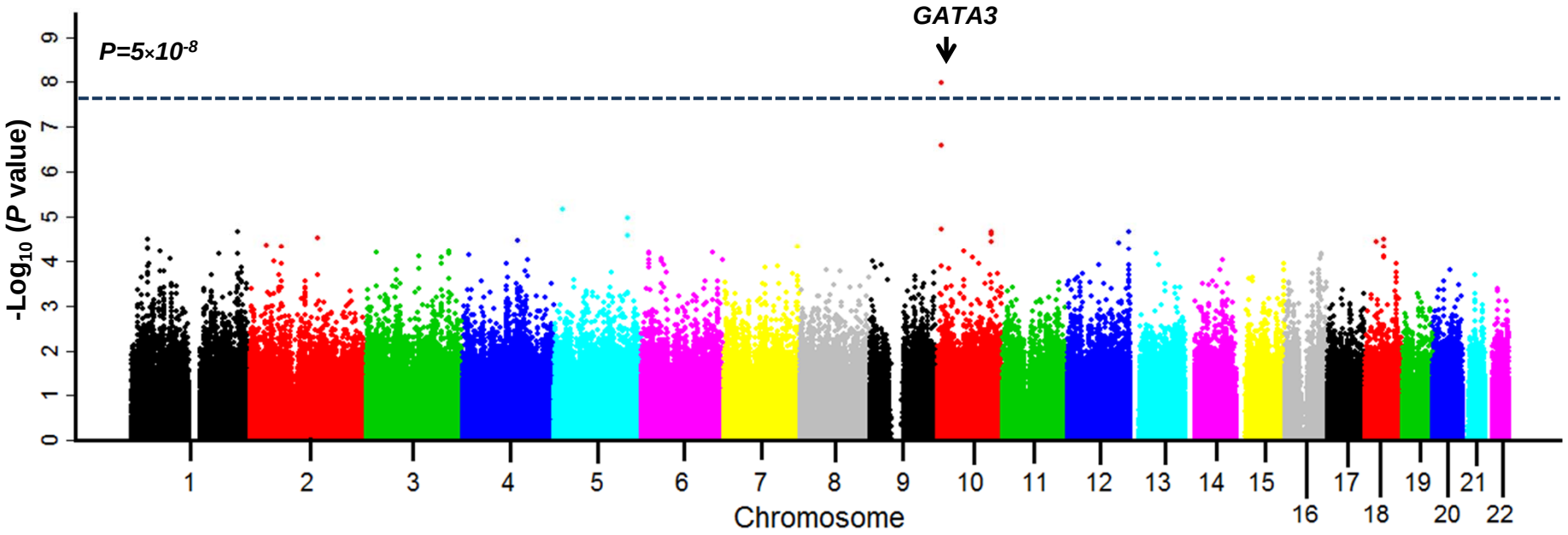
Supplementary Figure 1. Flow chart of SNP quality control/filtering in the discovery GWAS. SNPs were filtered on the basis of allele frequency and call rate, as detailed in "*Genotyping and Quality Control*".



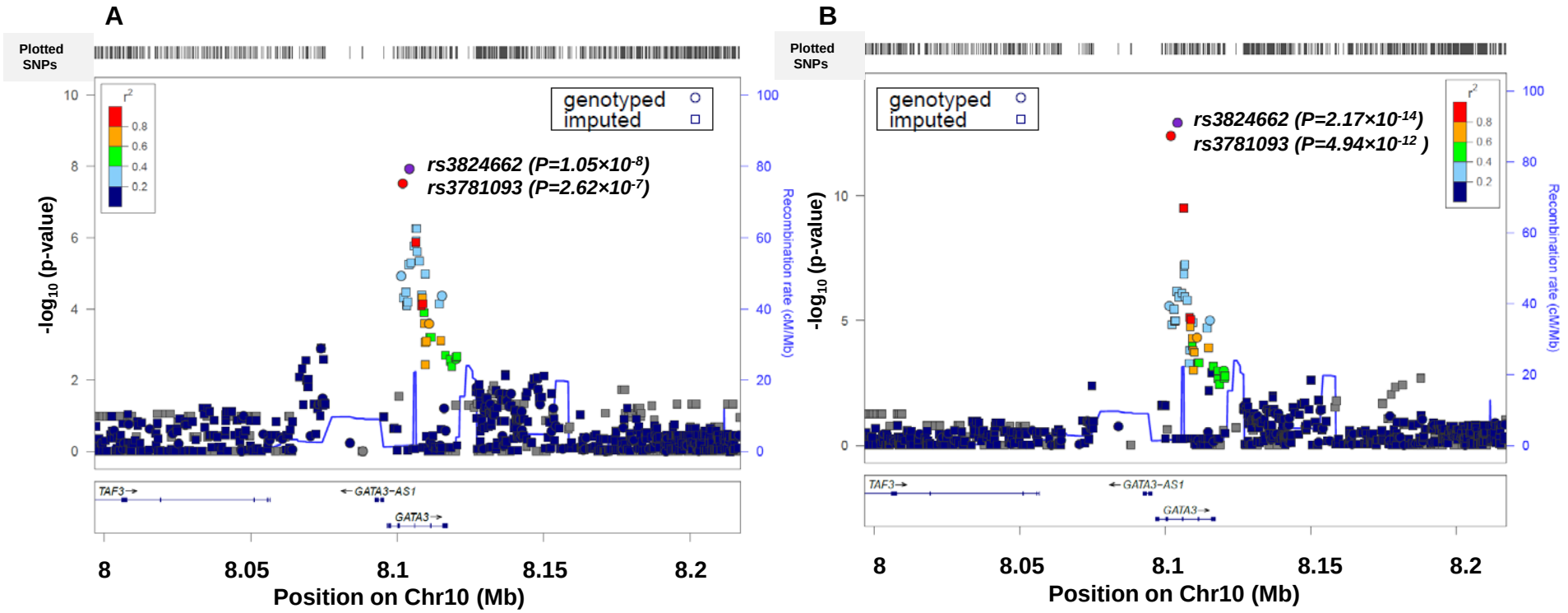
Supplementary Figure 2. Association results and linkage disequilibrium (LD) at the 10p14 locus. Panel A illustrates association signals in the discovery GWAS (Ph-like ALL *vs*. non-ALL controls). The negative logarithm of the *P* value (left axis) and recombination rate (right axis) are plotted for a 250 Kb window at the 10p14 locus, using LocusZoom (*Bioinformatics* 26: 2336). Color indicates LD (r^2) with rs3824662 in the HapMap CEU samples and chromosome position is based on hg18. In **panel B**, LD at this locus is depicted based on r^2 in HapMap CEU and MEX cell lines, and the plots were constructed using the HaploView software. *GATA3* SNPs (rs3824662 and rs3781093) were in high LD with each other (CEU: r^2 =0.94 and D'=1; MEX: r^2 =0.90 and D'=0.95) within LD block # 5 in both CEU and MEX populations.



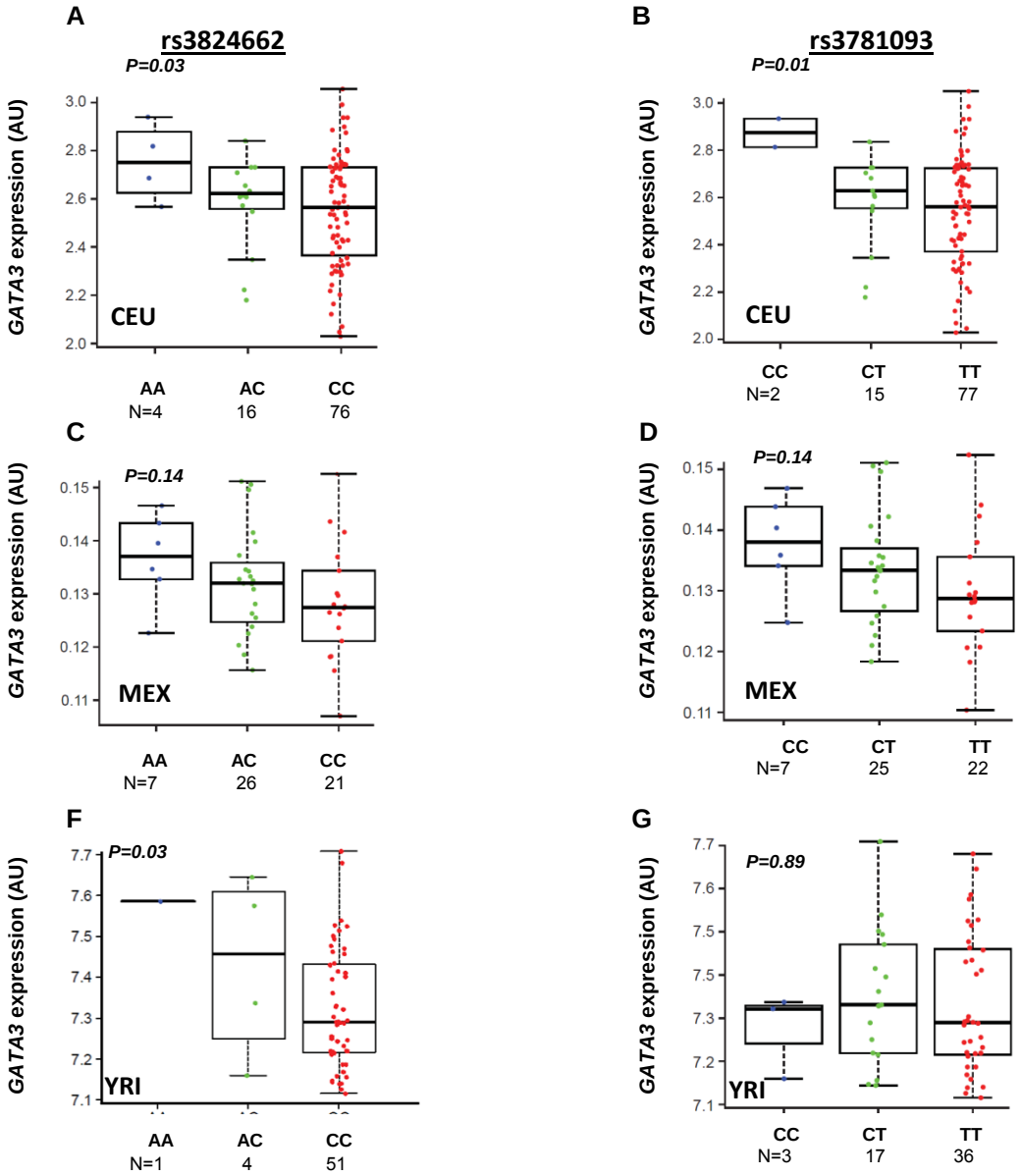
Supplementary Figure 3. Relationship between *GATA3* SNPs (rs3824662 and rs3781093) and Ph-like ALL by ethnicity in the COG AALL0232 cohort. The A allele at rs3824662 (**Panel A**) was over-represented in Ph-like ALL relative to non-Ph-like ALL and non-ALL controls. This association was true within the European Americans (>95% European genetic ancestry) or Hispanic Americans (>10% Native American genetic ancestry and Native American ancestry > African genetic ancestry). Similar association was confirmed for the risk allele (C) at the *GATA3* SNP rs3781093 (**Panel B**). Genetic ancestry was determined by using STRUCTURE (version 2.2.3) with HapMap CEU, YRI, CHB/JPT, and indigenous Native Americans as reference populations.



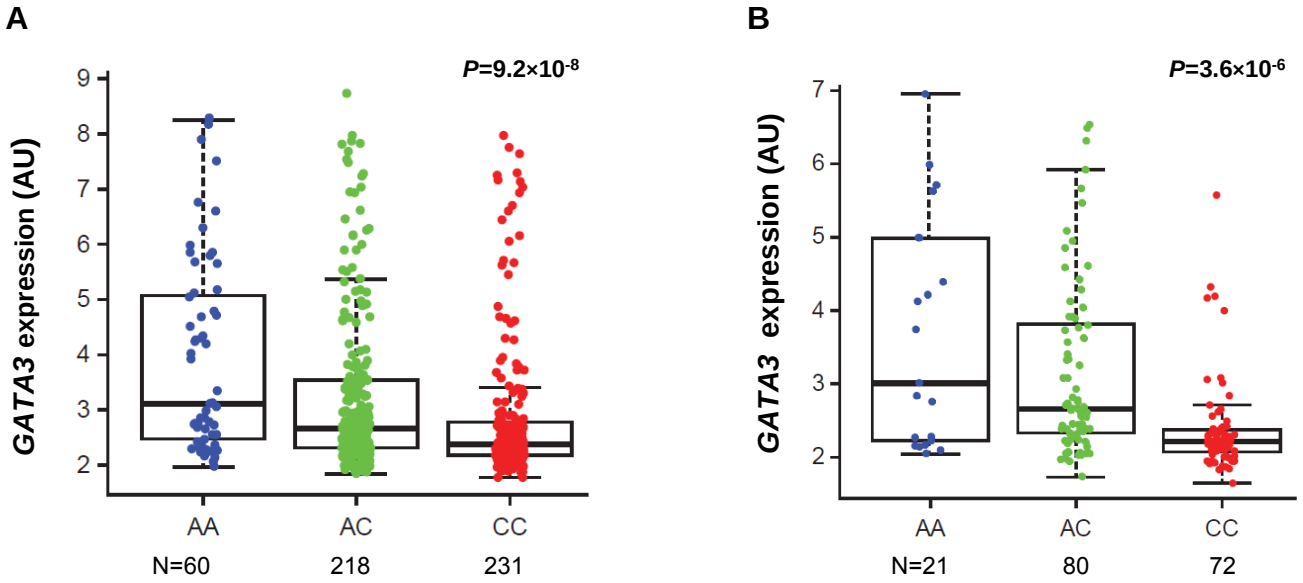
Supplementary Figure 4. GWAS of Ph-like ALL by comparing allele frequency between Ph-like ALL *vs.* non-Ph-like ALL. The association between genotype and Ph-like was evaluated using logistic regression model for 761,049 SNPs in 75 ALL cases with Ph-like gene expression profile and 436 ALL cases without this expression signature. *P*-values (-log 10 *P*, y axis) were plotted against respective chromosomal position of each SNP (x axis). Points above the blue horizontal line indicate SNPs achieving the genome-wide significant threshold ($P<5\times10^{-8}$). Gene symbol was indicated for the *GATA3* locus at 10p14.



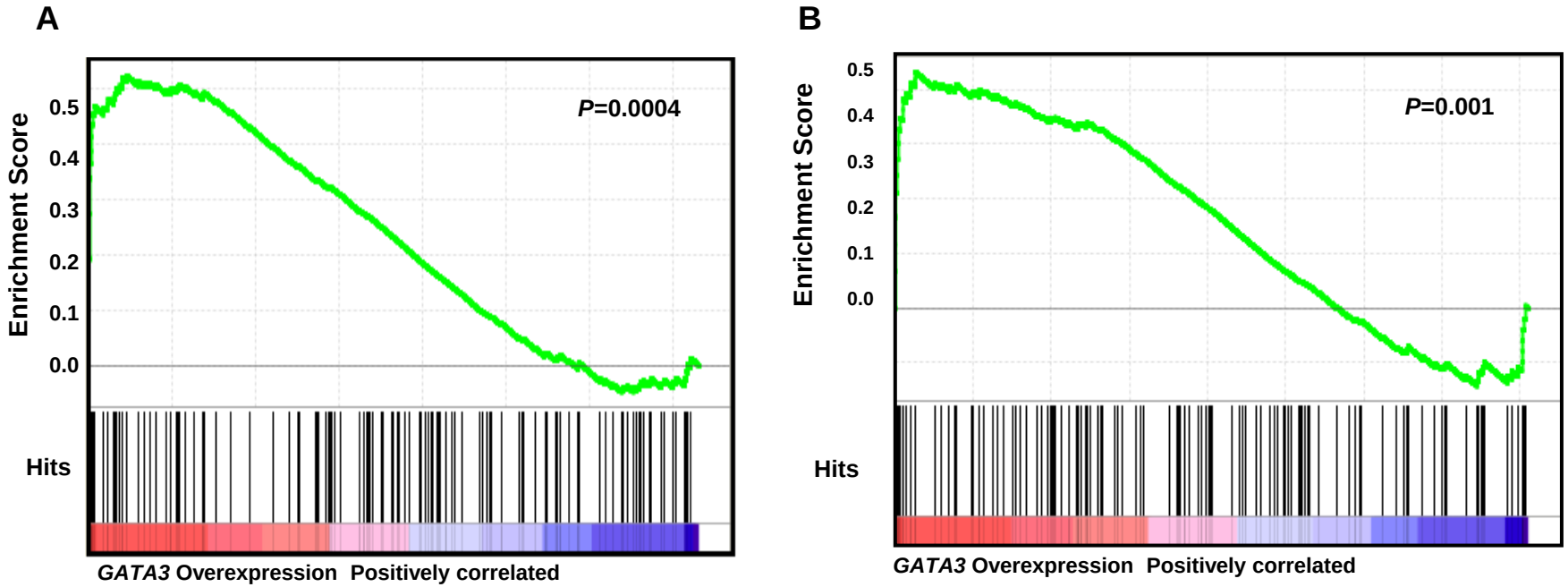
Supplementary Figure 5. Association with Ph-like ALL for imputed SNPs on chr10p14. Genotype was imputed for a 10 Mb region on 10p14 (chr10:60,523-10,060,447, hg19), using MaCH-Admix 2.0.185 with 1,000 Genome data set as references. Association was tested by comparing genotype frequency between Ph-like ALL (N=75) *vs.* non-Ph-like ALL (N=436, Panel **A**) and between Ph-like ALL (N=75) *vs.* non-ALL controls (N=6,661, Panel **B**) at 37,493 imputed or directly genotyped SNPs. Shown here are the association results for a 220 Kb window centered around rs3824662 with 727 SNPs spanning chr10:7,996,666-8,217,164 (hg19), and the plots are constructed using LocusZoom (*Bioinformatics* 26: 2336). Color indicates LD (r2) with rs3824662 in the HapMap CEU population.



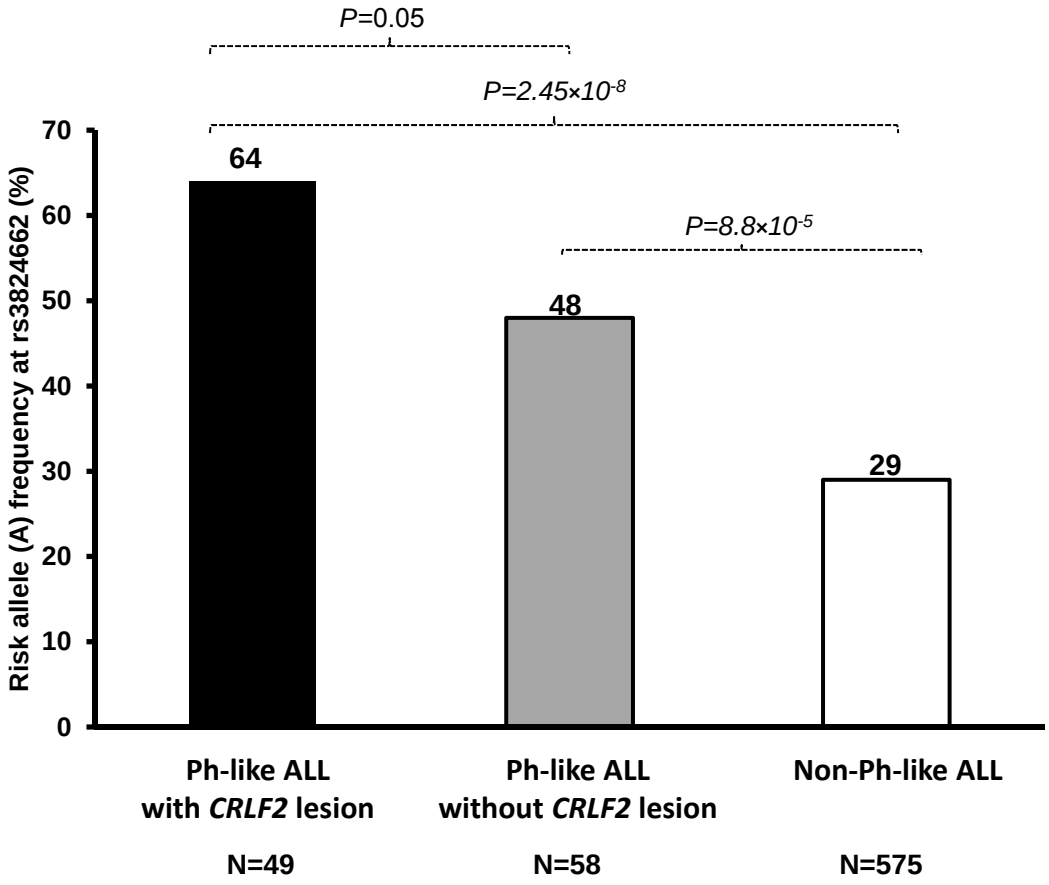
Supplementary Figure 6. **eQTL analyses of rs3824662 and rs3781093 in diverse HapMap populations.** *GATA3* SNP rs3824662 risk allele (A) and the rs3781093 (C) were associated with higher *GATA3* mRNA in 96 unrelated lymphoblastoid cell lines from the HapMap CEU population, using the publicly available gene expression data set GSE5859 (**A** and **B,** respectively). Similar trend was observed in in 54 unrelated HapMap MEX cell lines for which *GATA3* expression was evaluated by real time-PCR (Panels **C** and **D** for rs3824662 and rs3781093, respectively). Genotype-expression association in 56 unrelated YRI samples is represented in panel **F** (rs3824662) and panel **G** (rs3781093), using gene expression data set GSE7851. Genotype-expression association was evaluated using a linear regression model adjusting for ancestry as appropriate. AU, arbitrary units. Boxes include data between the twenty-fifth and the seventy-fifth percentiles.



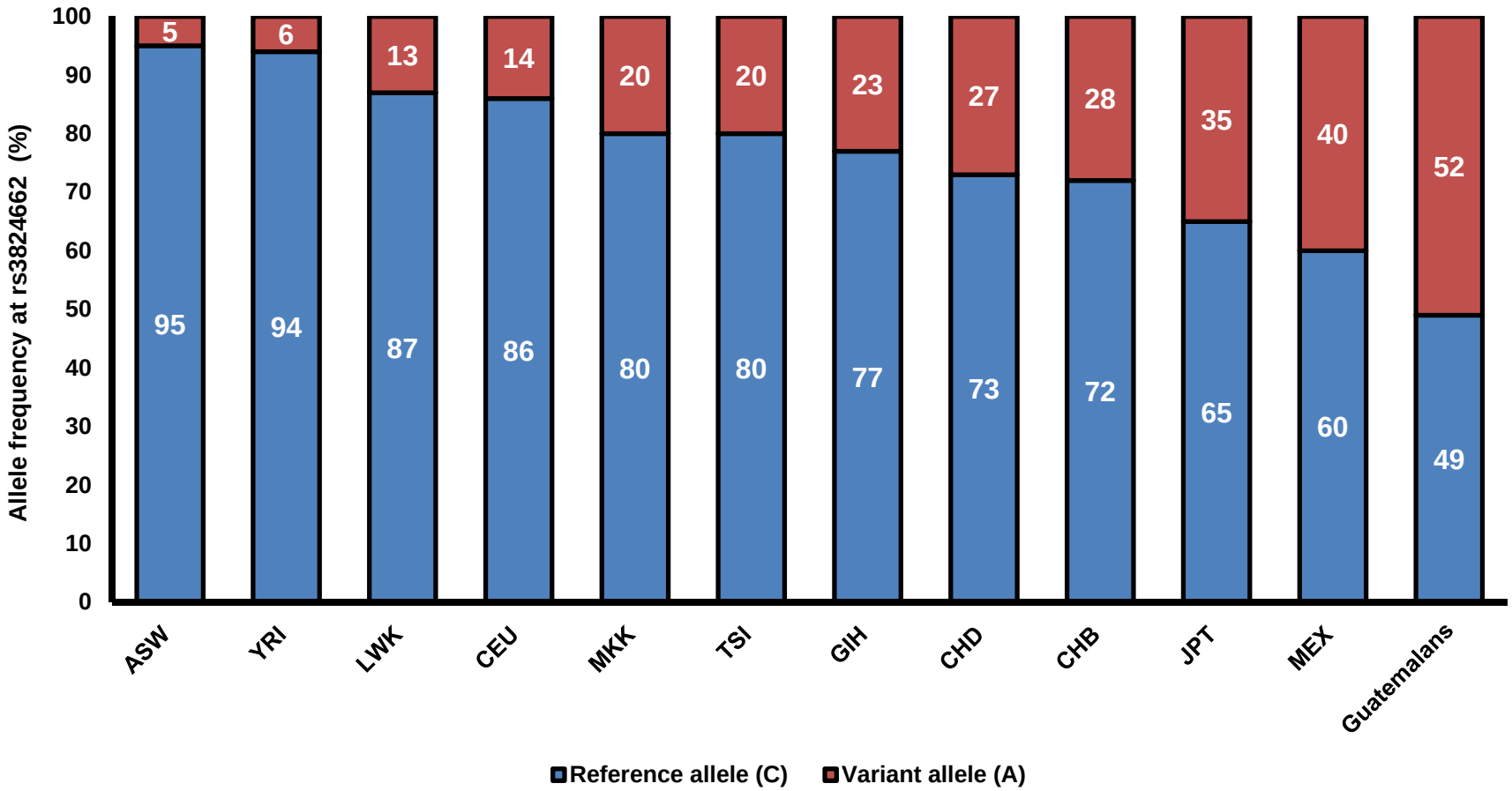
Supplementary Figure 7. rs3824662 genotype was associated with *GATA3* expression in ALL blasts. *GATA3* SNP rs3824662 risk allele (the A allele) was associated with higher *GATA3* mRNA expression in diagnostic ALL blasts from 511 children in the COG AALL0232 cohort (**A**) and 173 children in the COG P9906 cohort (**B**).Genotype-expression association was evaluated using a linear regression model, adjusting genetic ancestry as appropriate. AU, arbitrary unit. Boxes include data between the twenty-fifth and the seventy-fifth percentiles.



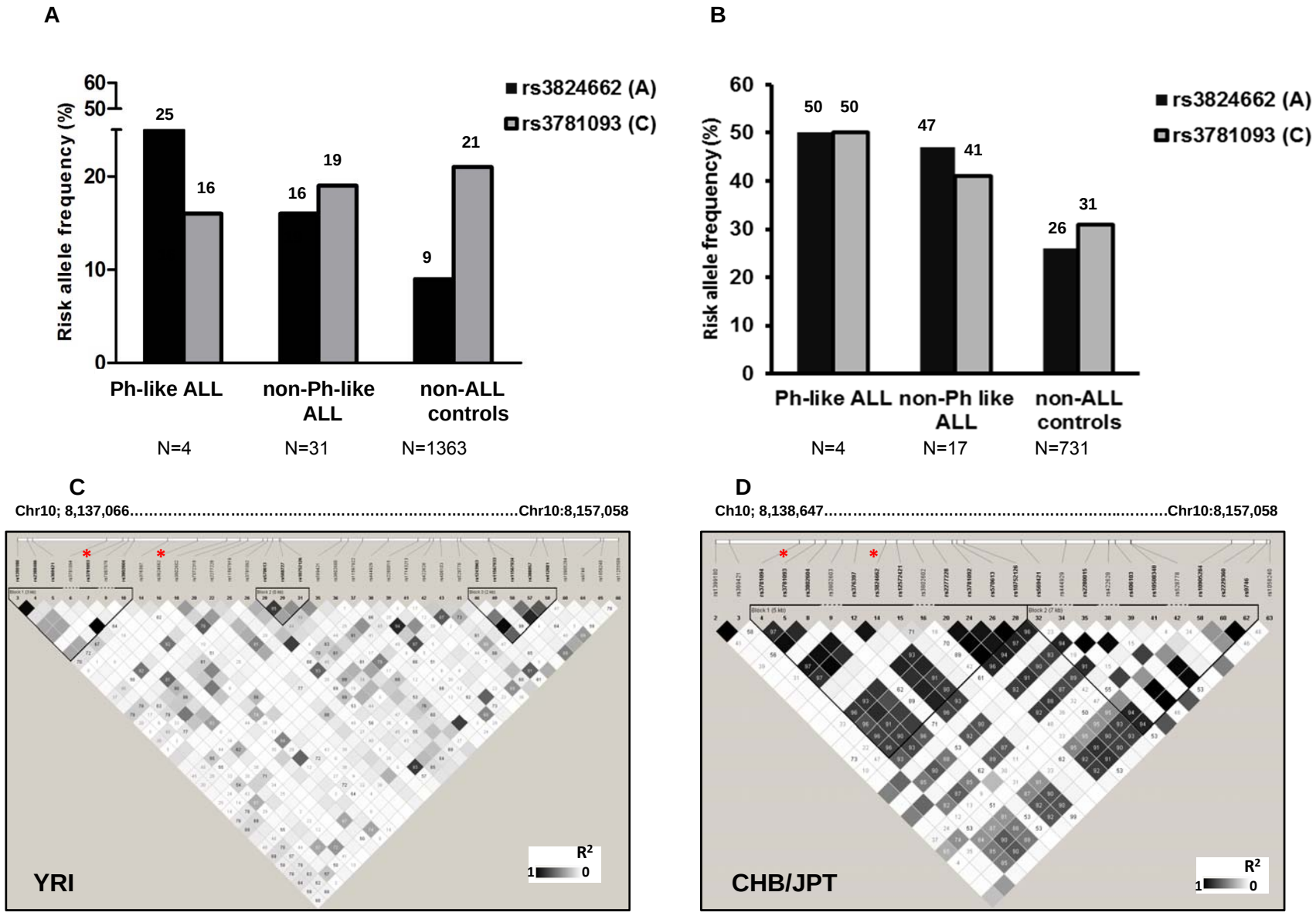
Supplementary Figure 8. Enrichment of Ph-like signature genes in ALL cell lines ectopically overexpressing *GATA3* compared with those transduced with control vectors, using the Gene Set Enrichment Analysis (GSEA). GSEA tested the upregulation of Ph-like ALL genes (i.e., PAM-based Ph-like signature) in ALL cell line UOCB1 (**A**) and Nalm6 (**B**) after ectopic overexpression of *GATA3* gene. *P* value was based permutations.



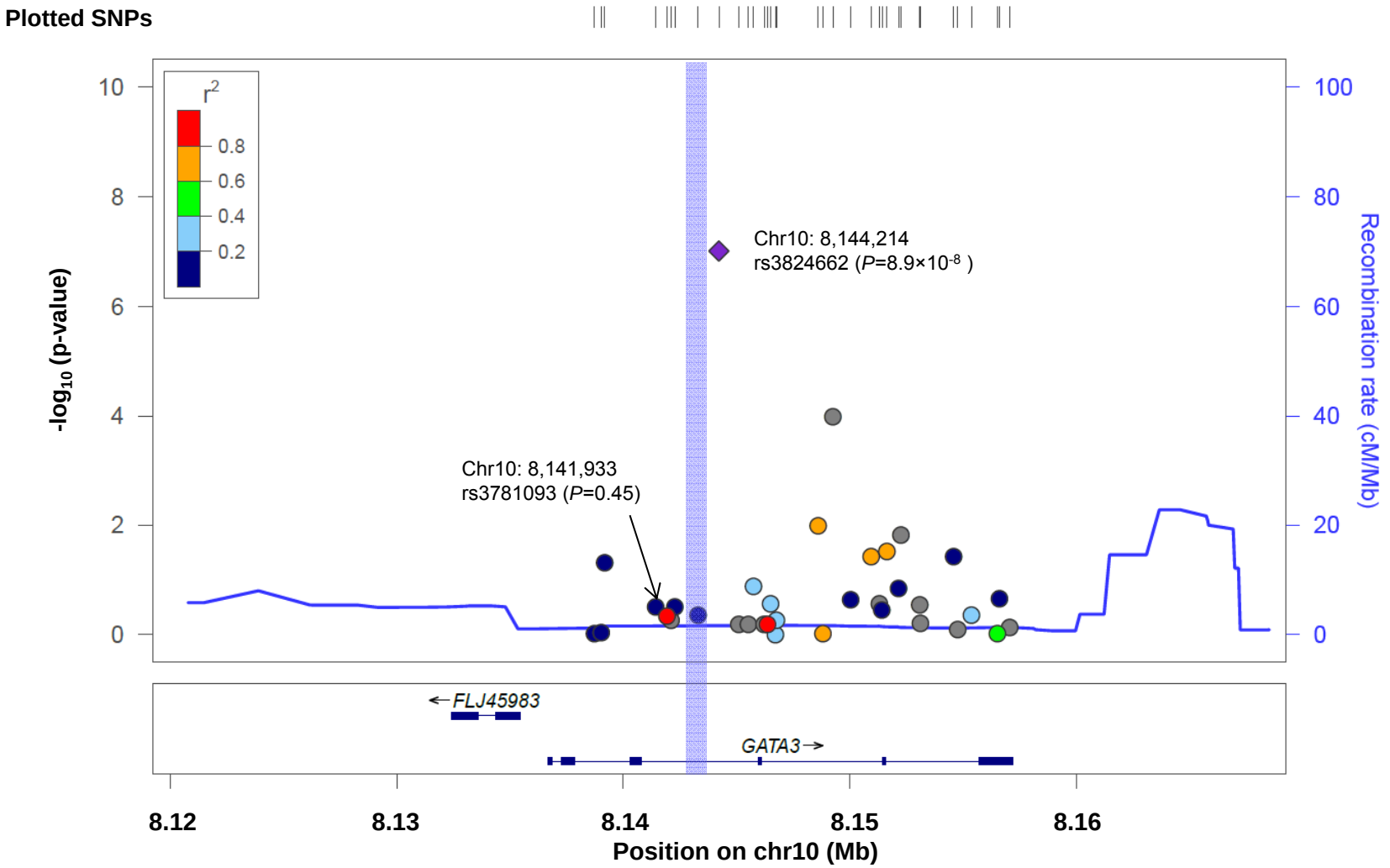
Supplementary Figure 9. Risk allele frequency at *GATA3* SNP rs3824662 in Ph-like patients (according to *CRLF2* status) and non-Ph-like patients. Combined cohort includes COG AALL0232 and COG P9906 (N=682), and *P* values were estimated by logistic regression test after adjusting for genetic ancestry.



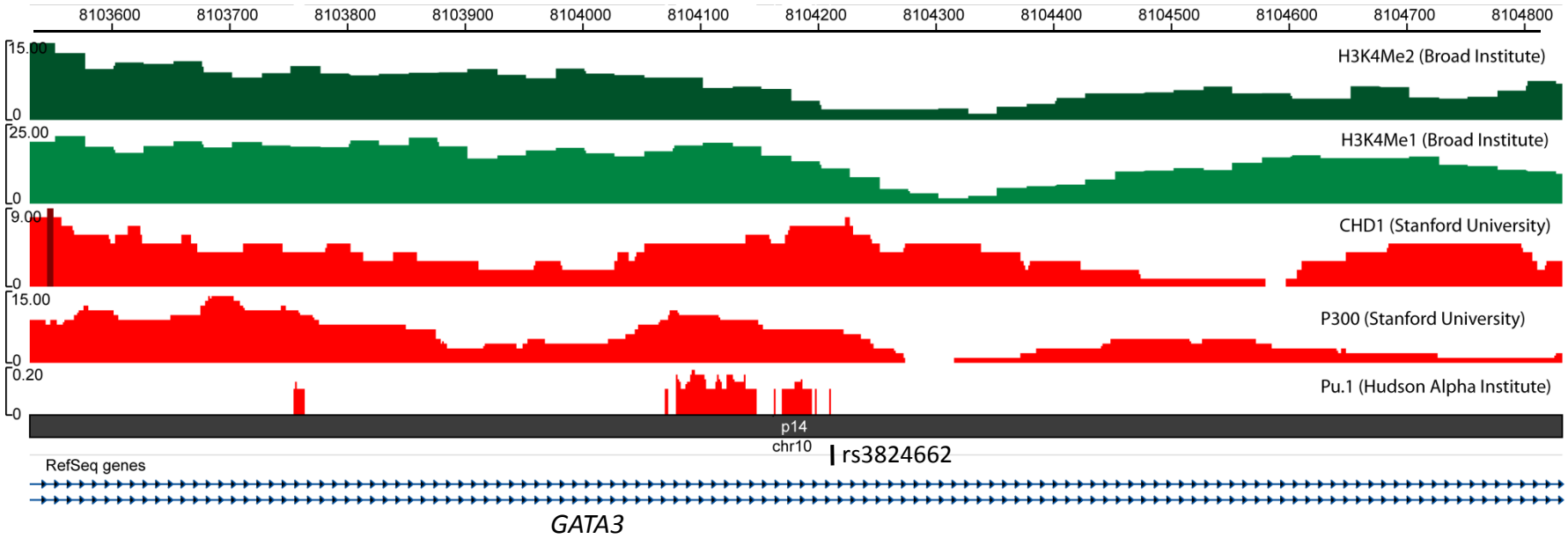
Supplementary Figure 10. *GATA3* SNPs rs3824662 allele frequency in worldwide populations. The frequency of the Ph-like ALL-related allele (A) is shown in an ascending order for HapMap populations and Native Americans in Guatemala. **Population descriptors: ASW**: African ancestry in Southwest USA, **CEU**: Utah residents with Northern and Western European ancestry from the CEPH collection, **CHB**: Han Chinese in Beijing, China, **CHD**: Chinese in Metropolitan Denver, Colorado, **GIH**: Gujarati Indians in Houston, Texas, **JPT**: Japanese in Tokyo, Japan, **LWK**: Luhya in Webuye, Kenya, **MEX**: Mexican ancestry in Los Angeles, California, **MKK**: Maasai in Kinyawa, Kenya, **TSI**: Tuscan in Italy, **YRI**: Yoruban in Ibadan, Nigeria. **Guatemalan**: 65 unrelated Guatemalan individuals with high Native American ancestry.



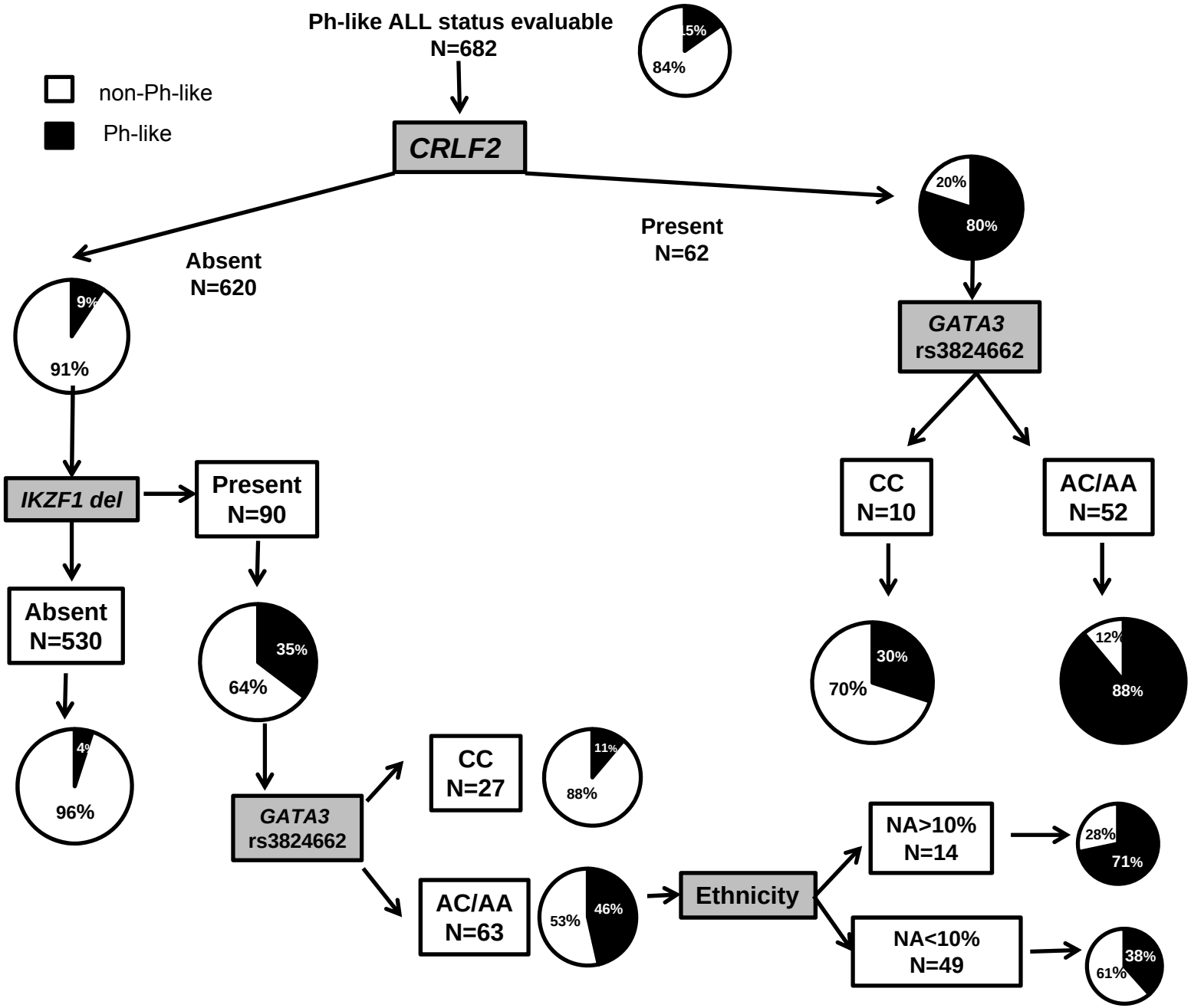
Supplementary Figure 11. **Association of *GATA3* SNPs with Ph-like ALL in African and Asian populations**. Panels **A** and **B** illustrate the allele frequency at rs3824662 and rs3781093 for Ph-like ALL, non-Ph-like ALL, and non-ALL controls with >70% African ancestry (**A**) and those with >90% Asian ancestry (**B**). In panels **C** and **D**, LD is depicted based on r^2 in unrelated HapMap YRI (**C**) and CHB/JPT (**D**) samples, and the plots were constructed using the HaploView software. rs3824662 and rs3781093 are in high LD with each other in CHB/JPT population (r^2 =0.97 and D'=1.0) but not in YRI population (r^2 =0.006 and D'=0.16). Chromosome position is based on hg18.



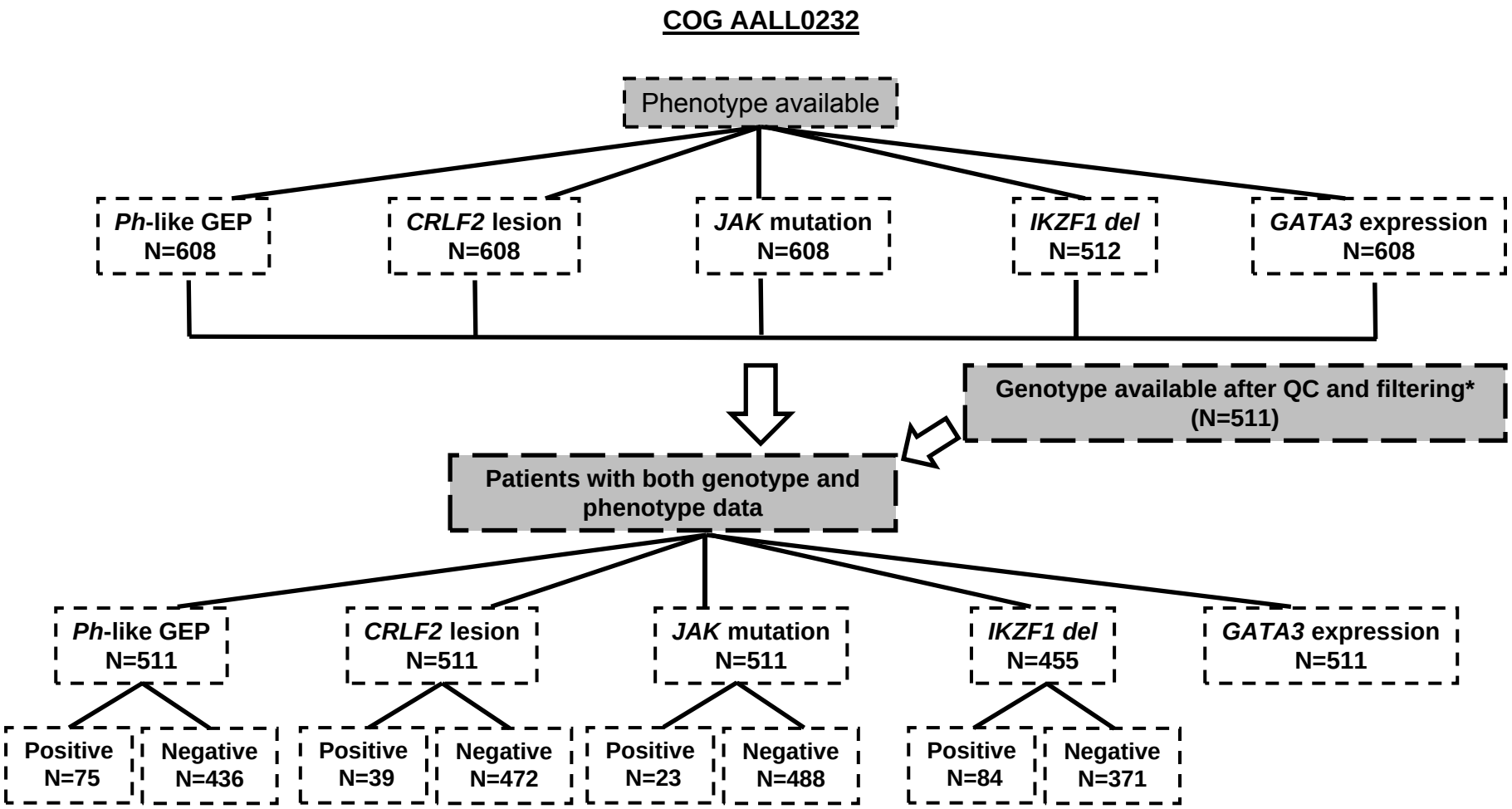
Supplementary Figure 12. Association of *GATA3* variants with local DNase hypersensitivity. DNase hypersensitivity was obtained for 70 HapMap YRI samples from a previously published data set (Nature 482:390), and genotype at 35 SNPs was retrieved from the 1,000 Genomes data set. rs3824662 (ch10; 8,144,214, hg18) showed the strongest association with the local DNase sensitivity window (chr10 8,144,000-8,144,100; shaded region), as determined by a linear regression test.



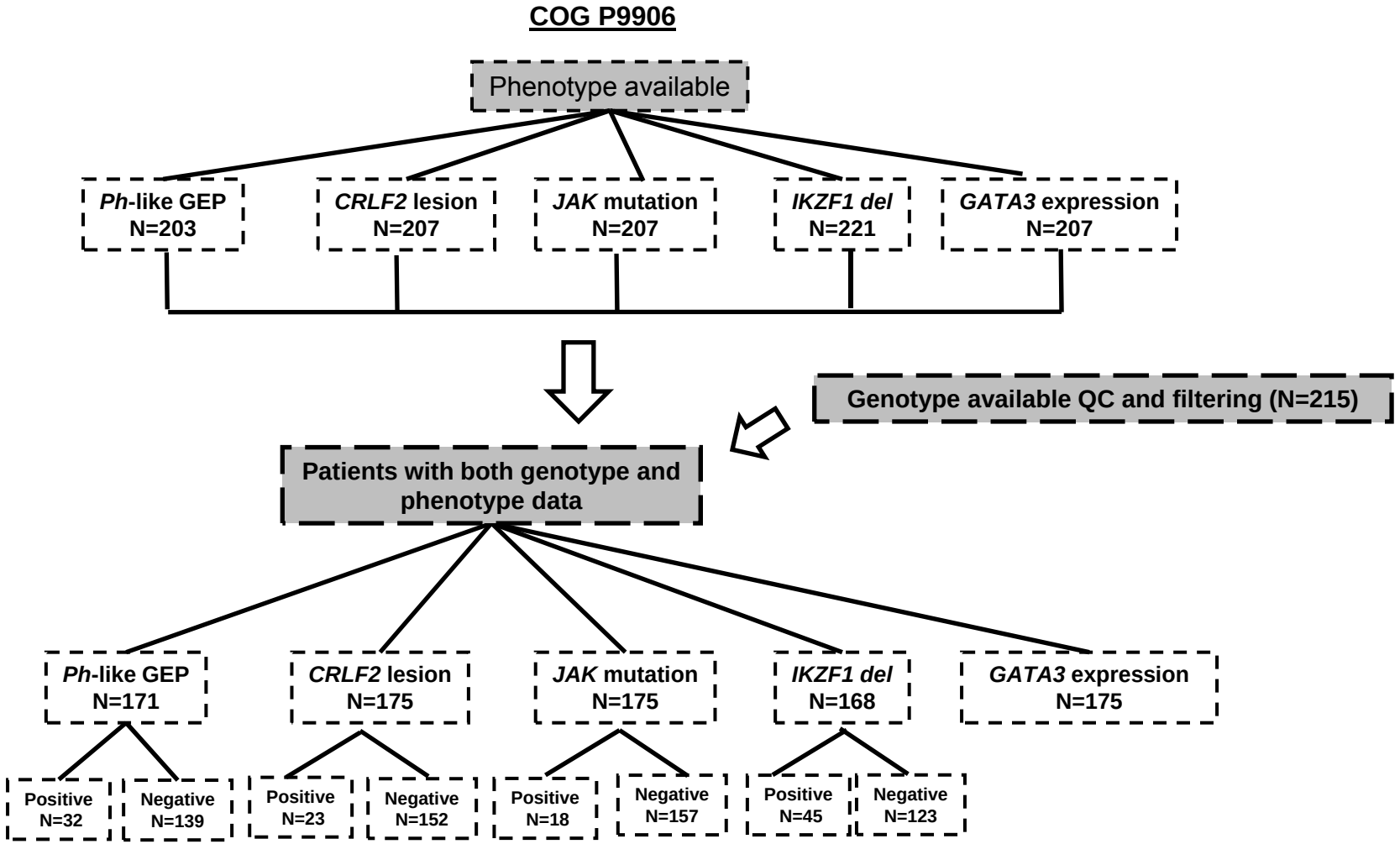
Supplementary Figure 13. Enhancer signal within the genomic region encompassing rs3824662 in the ENCODE data set. ENCODE data was queried for possible regulatory activities around rs3824662, focusing on histone marks and transcription factor binding. In GM12878 (ENCODE Tier 1), H3K4Me1 and H3K4Me2 signals indicate enhancer activity at rs3824662 and the reduction of ChIP-seq reads correspond to possible occupancy by transcription factor (P300 and PU.1). Consistent enrichment of CHD1-binding signal (chromatin remodeling activity) was also noted, plausibly in close proximity to the methylated H3K4. Graph is constructed using WashU Epigenome Browser (Nat Methods 10:375).



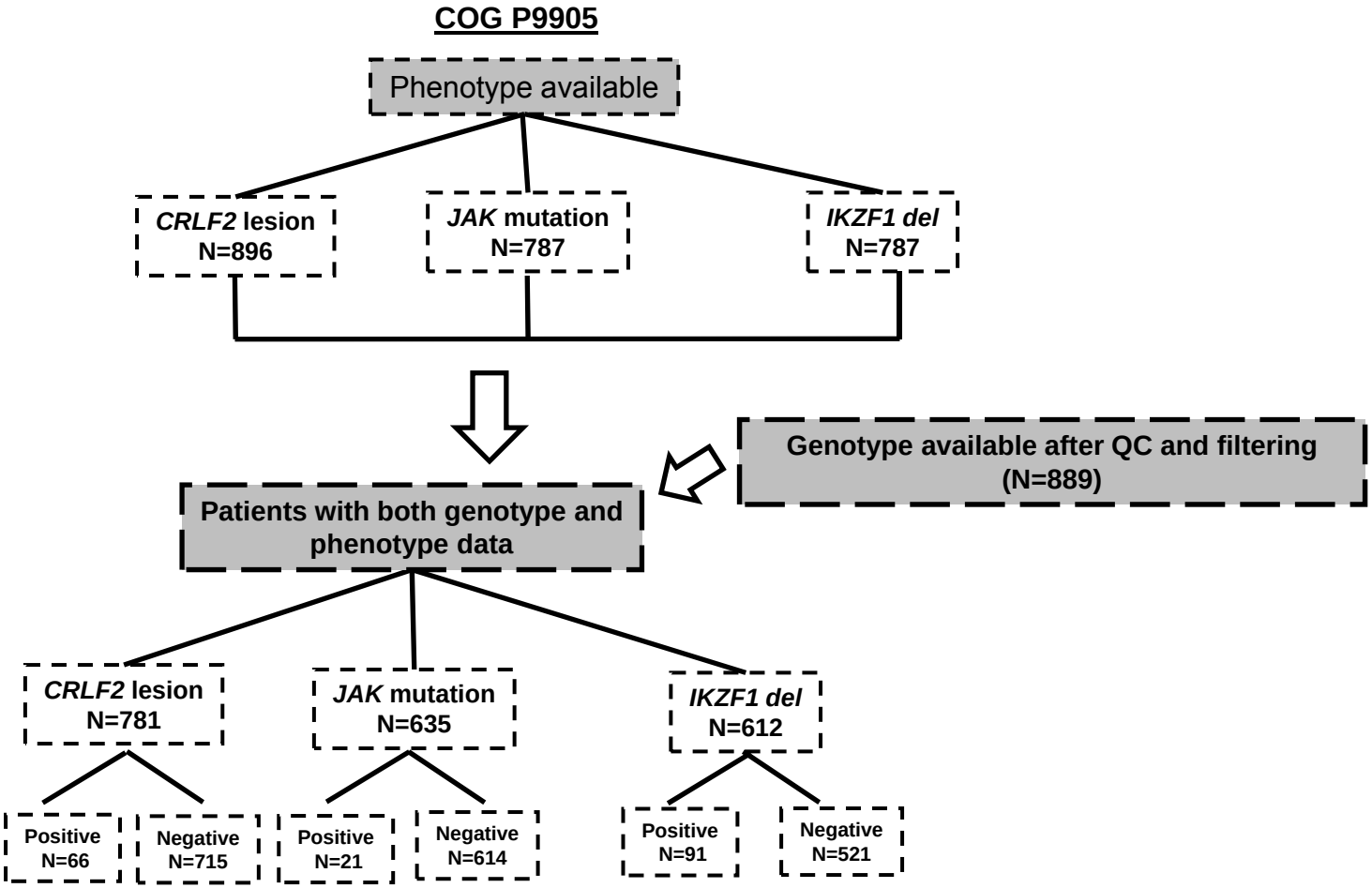
Supplementary Figure 14. Classification and regression tree (CART) analysis of predictors for Ph-like ALL. 682 patients with Ph-like ALL status evaluable from the COG AALL0232 and COG P9906 protocols were included. CART analysis was performed using the rpart function in R software, with *CRLF2*, *JAK*, *IKZF1* lesion, *GATA3* SNP, and genetic ancestry included in the model building process. NA, Native American genetic ancestry.



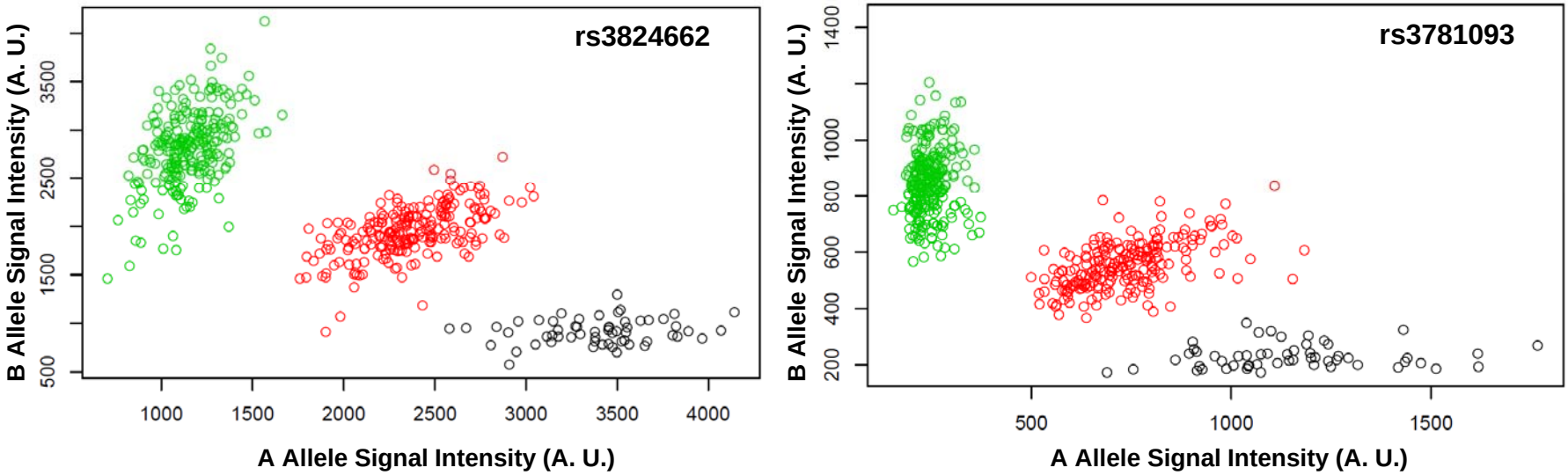
Supplementary Figure 15. Detailed description of patients tested for Ph-like, somatic *IKZF1* deletion, *CRLF2* lesion, *JAK* mutation in COG AALL0232 discovery GWAS group. *Of 550 sample genotyped, 12 samples were removed due to poor call rate or mismatch, and 27 Ph+ ALL were included in a separate analysis.



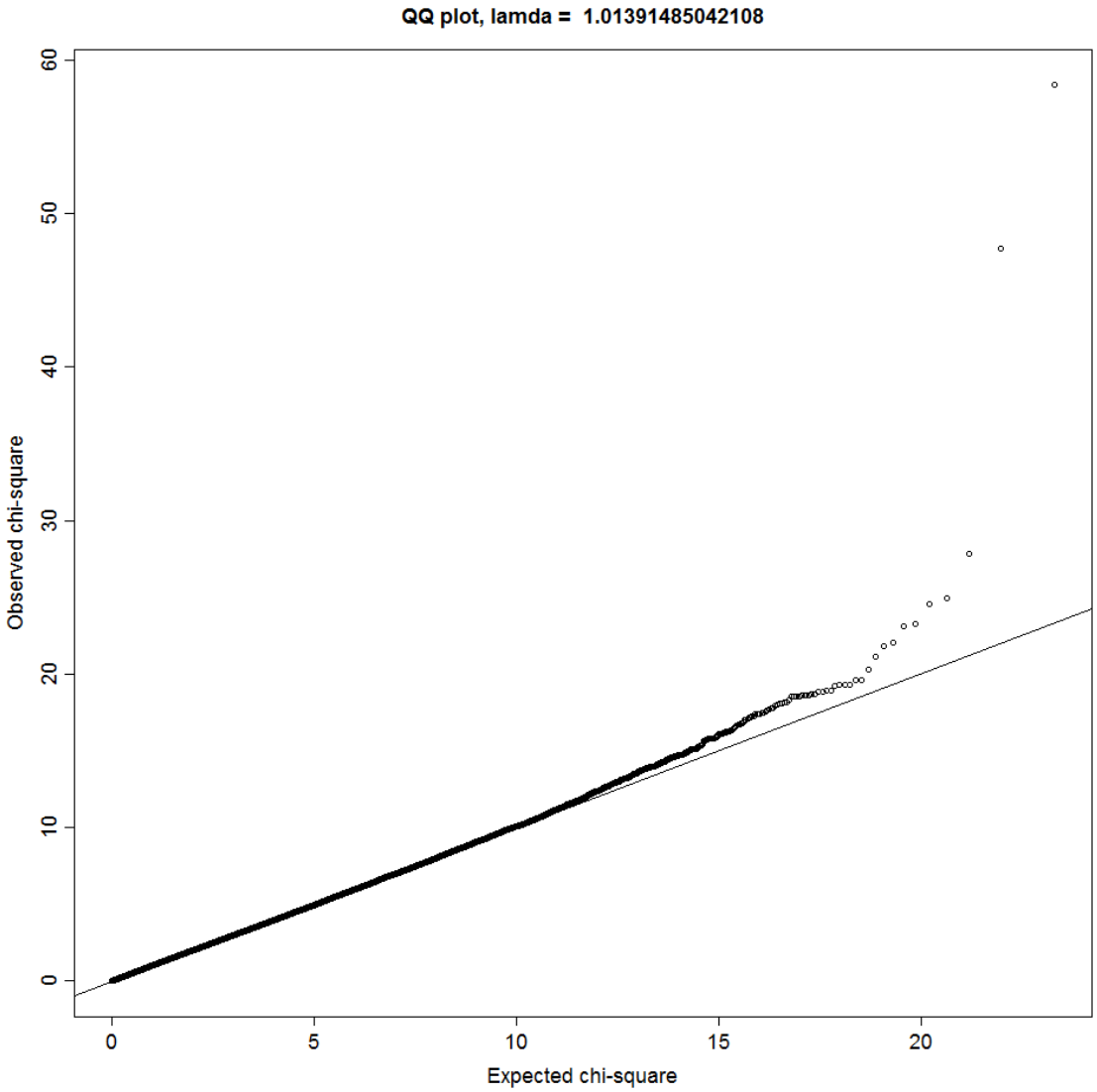
Supplementary Figure 16. Detailed description of patients tested for Ph-like, somatic *IKZF1* deletion, *CRLF2* lesion, *JAK* mutation in COG P9906 replication group.



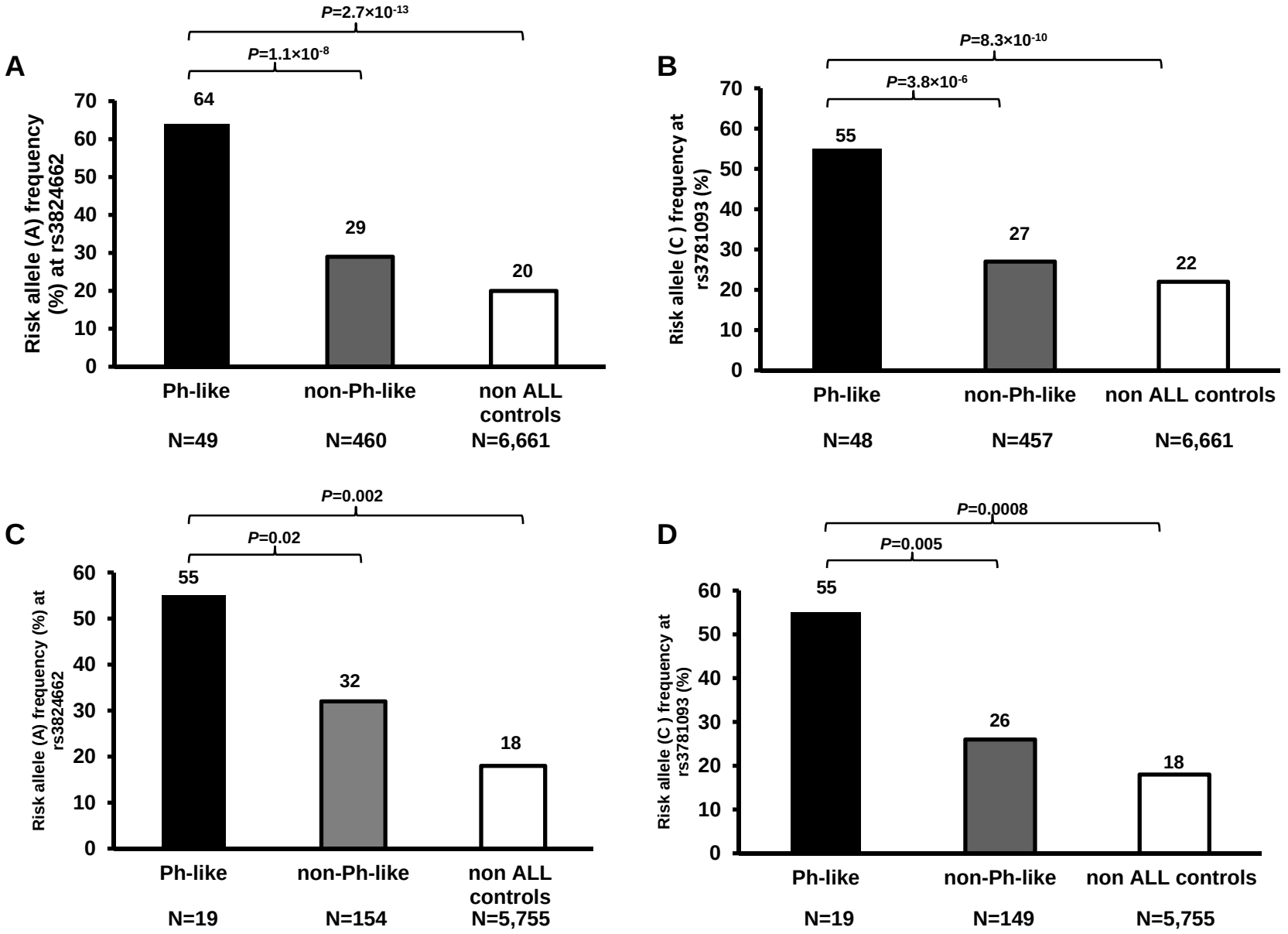
Supplementary Figure 17. Detailed description of patients tested for Ph-like, somatic *IKZF1* deletion, *CRLF2* lesion, *JAK* mutation in COG P9905 group.



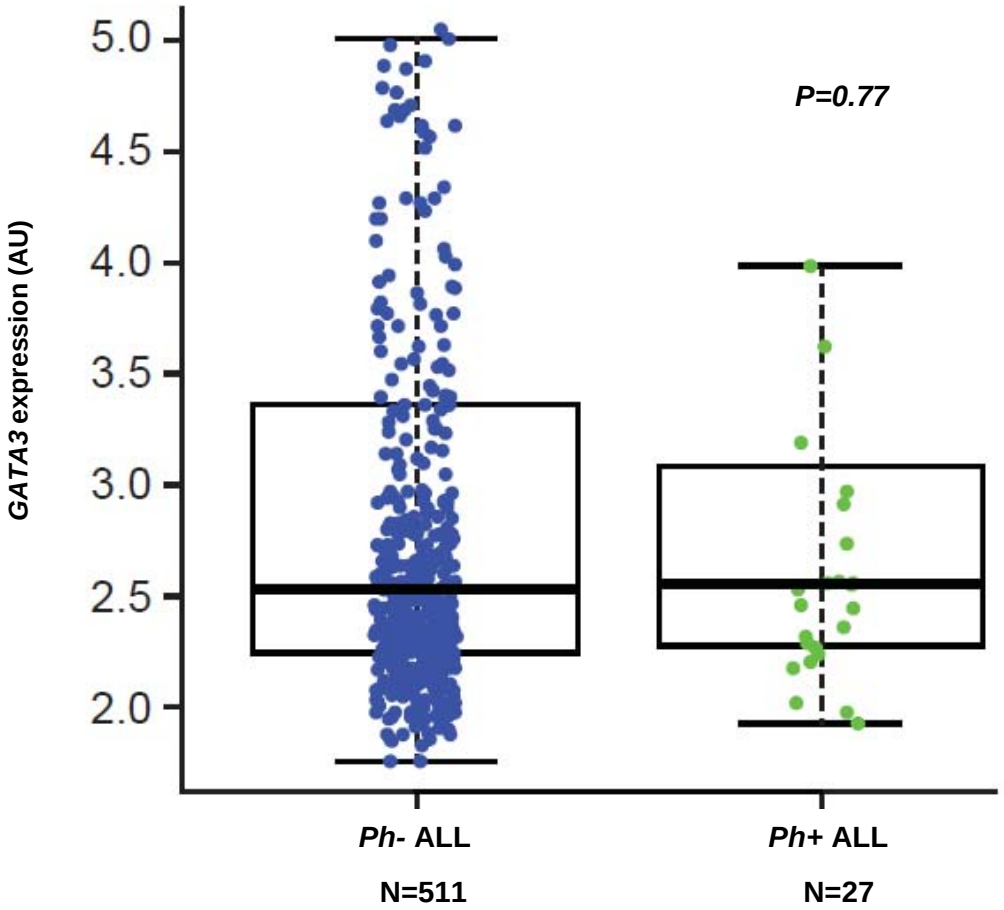
Supplementary Figure 18. Cluster plots of rs3824662 and rs3781093. Signal intensity for A and B alleles at both SNPs was based on theta value of 538 samples in the COG AALL0232 cohort, using Affymetrix Genotyping Console. At both SNPs, samples with AA, AB, and BB genotype clearly clustered into distinct groups, indicating high-quality genotype calls. AU, arbitrary units.



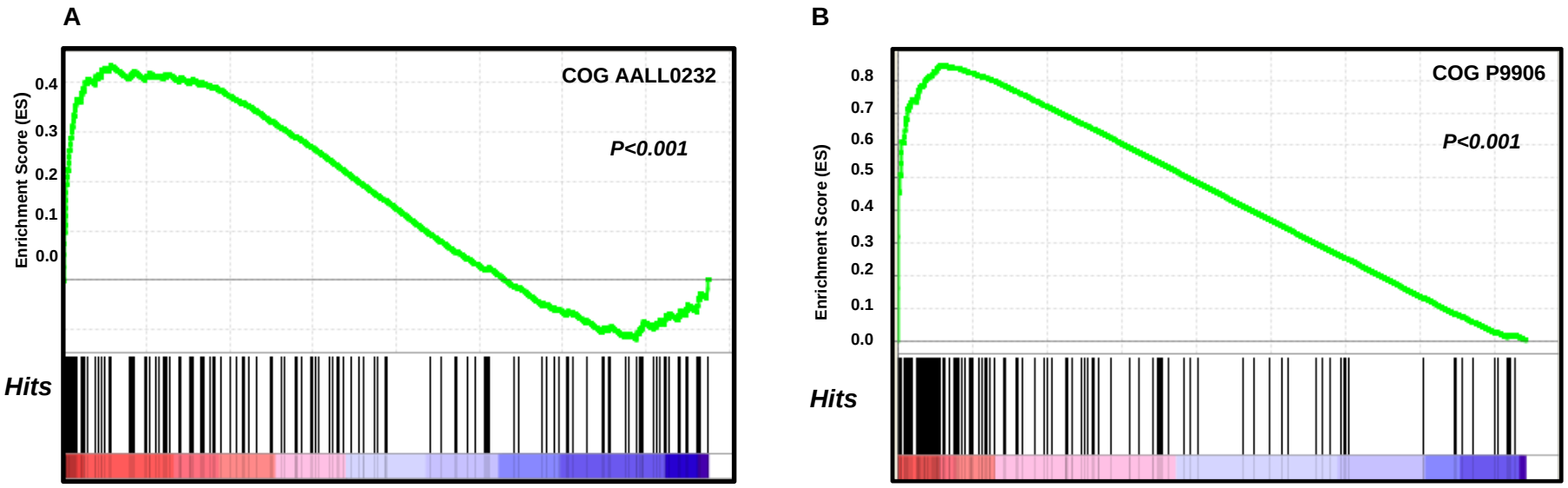
Supplementary Figure 19. Quantile-quantile (Q-Q) plot of logistic regression test for GWAS. The negative logarithm of the observed (y axis) and the expected (x axis) *P* value is plotted for each SNP (dot), and the black line indicates the null hypothesis of no true association. Deviation from the expected *P* value distribution is evident only in the tail area (λ=1.01), suggesting that population stratification was adequately controlled by adjusting for genetic ancestry.



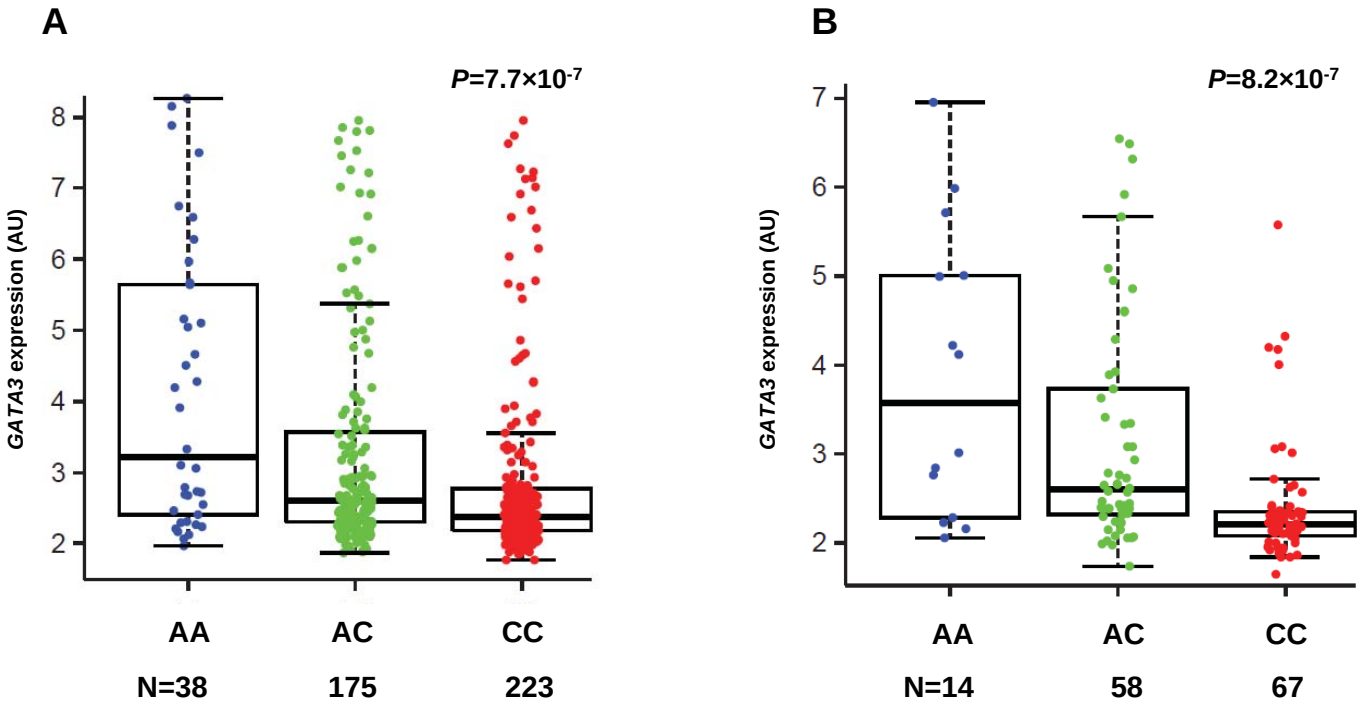
Supplementary Figure 20. Association between *GATA3* SNPs and Ph-like ALL identified on the basis of ROSE clustering. *GATA3* SNPs rs3824662 (A) and rs3781093 (B) were associated with Ph-like ALL in the discovery GWAS group (COG AALL0232 and dbGAP-MESA). The association was also validated in the replication group (COG P9906 and independent non-ALL controls, Panels C and D). *P*-values were estimated by the logistic regression test after adjusting for ancestry.



Supplementary Figure 21. ***GATA3* expression in Ph-positive ALL *vs.* Ph-negative ALL.** *GATA3* expression was quantified by Affymetrix U133A array in diagnostic bone marrow in COG AALL0232 and association with Ph+ status was tested using a logistic regression model adjusting for ancestry as appropriate. AU, arbitrary units. Boxes include data between the twenty-fifth and the seventy-fifth percentiles.



Supplementary Figure 22. Enrichment of Ph-like signature in genes differentially expressed by rs3824662 genotype in ALL blasts. We first identified genes for which expression was associated with *GATA3* rs3824662 genotype (AA+AC *vs.* CC) in COG AALL0232 and COG P9906 cohorts. Over-representation of the Ph-like gene signature in those affected by *GATA3* SNP genotype was then evaluated using GSEA and *P* value was estimated based on permutations.



Supplementary Figure 23. **rs3824662 was associated with *GATA3* expression in non-Ph-like ALL.** The A allele at rs3824662 was associated with higher *GATA3* expression in non-Ph-like ALL cases in the COG AALL0232 (N=436; $P=7.7\times10^{-7}$;**Panel A**) and COG P9906 cohorts (N=139; $P=8.2\times10^{-7}$; **Panel B**), indicating direct influence of SNP on *GATA3* transcription. Genotype-expression association was evaluated using a linear regression model adjusting by ancestry as appropriate. AU, arbitrary units. Boxes include data between the twenty-fifth and the seventy-fifth percentiles.