

Title	Clustering high-dimensional mixed data to uncover sub- phenotypes: joint analysis of phenotypic and genotypic data
Authors	McParland, D.;Phillips, Catherine M.;Brennan, L.;Roche, H. M.;Gormley, I. C.
Publication date	2017-06-30
Original Citation	McParland, D., Phillips, C. M., Brennan, L., Roche, H. M. and Gormley, I. C. (2017) 'Clustering high-dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data', Statistics in Medicine, 36(28), pp. 4548-4569. doi:10.1002/sim.7371
Type of publication	Article (peer-reviewed)
Link to publisher's version	10.1002/sim.7371
Rights	© 2017, John Wiley & Sons, Ltd. This is the peer reviewed version of the following article: McParland, D., Phillips, C. M., Brennan, L., Roche, H. M. and Gormley, I. C. (2017) 'Clustering high- dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data', Statistics in Medicine, 36(28), pp. 4548-4569. doi:10.1002/sim.7371, which has been published in final form at https://doi.org/10.1002/sim.7371. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.
Download date	2025-01-28 16:10:32
Item downloaded from	https://hdl.handle.net/10468/6192



University College Cork, Ireland Coláiste na hOllscoile Corcaigh

# Clustering high dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data. McParland, D., Phillips, C.M., Brennan, L., Roche, H.M. and Gormley, I.C. Supplementary material.

### 1 Complete list of variables collected

Apo A-1 (g/L)	Apo B (g/L)	Body Mass Index (kg/m2)
Cholesterol (mmol/L)	Diastolic Blood Pressure (mm Hg)	Glucose $(mmol/L)$
plasma.C14.0 (*)	plasma.C16.0 (*)	plasma.C16.1 (*)
plasma.C18.0 $(*)$	plasma.C18.1 (*)	plasma.c18.2 (*)
plasma.C18.3.n.6 (*)	plasma.C18.3.n.3 (*)	plasma.C18.4 (*)
plasma.C20.1 (*)	plasma.C20.3 (*)	plasma.C20.4.n.6~(*)
plasma.C20.4.n.3 (*)	plasma.C20.5 (*)	plasma.C22.4 (*)
plasma.C22.5 (*)	plasma.C22.6 (*)	Systolic Blood Pressure (mm Hg)
Triglycerides (mmol/L)	Waist (cm)	

Table 1: Complete list of the 26 continuous phenotypic variables collected. Units are detailed in parentheses, where (\*) denotes percentage of the total quantified fatty acids.

<i>ABCA1</i> rs363717	<i>ABCA1</i> rs2066881	<i>ABCA1</i> rs2230808	<i>ABCA1</i> rs2297404
<i>ABCA1</i> rs4149313	<i>ABCA1</i> rs1929842	<i>ABCA1</i> rs2515616	<i>ABCA1</i> rs2791952
<i>ABCA1</i> rs2472510	<i>ABCA12</i> rs17430358	<i>ABCA12</i> rs10932587	<i>ABCA12</i> rs2970968
ABCC8 rs1799859	<i>ABCC8</i> rs2073583	ABCC8 rs916829	<i>ABCC9</i> rs829060
<i>ABCD1</i> rs5987140	<i>ABCD1</i> rs6643643	ABCD2 rs7968837	<i>ABCD2</i> rs10877201
ABCD2 rs4072006	<i>ABCD2</i> rs6581228	<i>ABCG</i> 4 rs3802885	ABCG5 rs6720173
ABCG5 rs4148189	ABCG5 rs4245786	ABCG5 rs11887534	ABCG8 rs6709904
ABCG8 rs4148217	ABCG8 rs6544718	ACACA rs725038	<i>ACACA</i> rs17138899
ACACB rs2300453	ACACB rs4766587	ACACB rs2284689	<i>ACACB</i> rs2075263
ACADL rs16844213	<i>ACADM</i> rs12032051	ACADM rs8763	ACADS rs3999408
ACADS rs695950	ACAT1 rs2280332	ACAT2 rs2146162	ADIPOQ rs266729
ADIPOQ rs17366568	ADIPOQ rs2241766	<i>ACSL1</i> rs4862417	ACSL1 rs9997745
ACSM2A rs1133607	ACSM2B rs16970280	ACSM2B rs7499304	<i>ADD1</i> rs12503220
<i>ADD1</i> rs6600769	<i>ADD1</i> rs4961	<i>ADD1</i> rs17777371	<i>ADD1</i> rs17834108

<i>ADD1</i> rs4963
ADRA2A rs3750625
AGER rs2070600
AGT rs3889728
<i>APOA2</i> rs5085
APOA4 rs5092
APOA5 rs3135506
<i>APOBEC1</i> rs2302515
APOC3 rs4520
APOH rs8178847
<i>APOL1</i> rs2012928
C3 rs344548
C3 rs2250656
GPR35 rs2975776
CEBPA rs12691
CPT2 rs1056438
CYP7A1 rs11786580
<i>ENPP1</i> rs1044498
ESR1 rs9340954
<i>FABP1</i> rs1530273
<i>FABP3</i> rs951545
FABP7 rs1564900
<i>FOXO1</i> rs7981045
GABRA6 rs6898571
GAD2 rs3781117
GCK rs12673242
GHRL rs696217
HK2 rs3771798
HNF4A rs2144908
<i>IDE</i> rs551266
<i>IGF1</i> rs6219
<i>IGF1</i> rs10860869
<i>PDX1</i> rs1124607
<i>PDX1</i> rs1124607
<i>KCNIP2</i> rs2305189
LDLR rs6511720
LEPR rs1137100
LIPC rs1800588
LIPG rs4245232
LOCIONEDWOOF ZADOOAO
LOC100507205 rs7480010
<i>LPIN1</i> rs11524
<i>LOC100507205</i> rs7480010 <i>LPIN1</i> rs11524 <i>LPL</i> rs1800590
LOCI00507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611
LOC100507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611 LRP2 rs4668123
LOC100507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611 LRP2 rs4668123 LRP5 rs587397
LOC100507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611 LRP2 rs4668123 LRP5 rs587397 NCOA1 rs2083389
LOCI00507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611 LRP2 rs4668123 LRP5 rs587397 NCOA1 rs2083389 NFATC3 rs7205935
LOC100507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611 LRP2 rs4668123 LRP5 rs587397 NCOA1 rs2083389 NFATC3 rs7205935 NOS3 rs743507
LOC100507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611 LRP2 rs4668123 LRP5 rs587397 NCOA1 rs2083389 NFATC3 rs7205935 NOS3 rs743507 NPY rs16478
LOC100507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611 LRP2 rs4668123 LRP5 rs587397 NCOA1 rs2083389 NFATC3 rs7205935 NOS3 rs743507 NPY rs16478 OLR1 rs1050289
LOC100507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611 LRP2 rs4668123 LRP5 rs587397 NCOA1 rs2083389 NFATC3 rs7205935 NOS3 rs743507 NPY rs16478 OLR1 rs1050289 PFDN1 rs9042
LOC100507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611 LRP2 rs4668123 LRP5 rs587397 NCOA1 rs2083389 NFATC3 rs7205935 NOS3 rs743507 NPY rs16478 OLR1 rs1050289 PFDN1 rs9042 PPARGC1A rs3736265
LOCI00507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611 LRP2 rs4668123 LRP5 rs587397 NCOA1 rs2083389 NFATC3 rs7205935 NOS3 rs743507 NPY rs16478 OLR1 rs1050289 PFDN1 rs9042 PPARGC1A rs3736265 PPARGC1A rs1878949
LOC100507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611 LRP2 rs4668123 LRP5 rs587397 NCOA1 rs2083389 NFATC3 rs7205935 NOS3 rs743507 NPY rs16478 OLR1 rs1050289 PFDN1 rs9042 PPARGC1A rs1878949 PPARGC1B rs7732671
LOC100507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611 LRP2 rs4668123 LRP5 rs587397 NCOA1 rs2083389 NFATC3 rs7205935 NOS3 rs743507 NPY rs16478 OLR1 rs1050289 PFDN1 rs9042 PPARGC1A rs3736265 PPARGC1A rs1878949 PPARGC1B rs7732671 PIK3R1 rs16897511
LOC100507205 rs7480010 LPIN1 rs11524 LPL rs1800590 LPL rs1059611 LRP2 rs4668123 LRP5 rs587397 NCOA1 rs2083389 NFATC3 rs7205935 NOS3 rs743507 NPY rs16478 OLR1 rs1050289 PFDN1 rs9042 PPARGC1A rs3736265 PPARGC1A rs1878949 PPARGC1B rs7732671 PIK3R1 rs16897511 PLIN1 rs8179037

*ADIPOR1* rs10753929 *ADRB3* rs4994 *AGRP* rs28937570 *AKT2* rs7247515 *APOA2* rs6413453 *APOA5* rs651821 *APOA5* rs662799 *APOBEC1* rs1015102 *APOC3* rs5128 *APOL1* rs2239785 *APOL6* rs5995133 *C3* rs8107911 *C3* rs163913 *CARTPT* rs3857384 *CEBPA* rs16967952 CREBBP rs130005 *ENPP1* rs10457576 *ESR1* rs11155814 *ESR1* rs2474148 *FABP2* rs1799883 *FABP4* rs1486006 *FOXC2* rs1035550 *FXN* rs1800651 *GAD2* rs8190582 *GAD2* rs11015008 *GCK* rs2908289 *GIPR* rs11671664 *HK2* rs651071 *HNF4A* rs3212180 *IDE* rs1887922 *IGF1* rs978458 *IGF2* rs680 *IGF2* rs3168310 *IRS1* rs1801278 *KCNIP2* rs550 LDLR rs8110695 *LEPR* rs10493380 *LIPC* rs2070895 *LIPG* rs3826577 *LPIN1* rs2577262 *LPIN1* rs4669781 LPL rs270 LPL rs268 *LRP2* rs16823023 MKKS rs1547 *NCOA1* rs11125744 *NFATC3* rs12598 *NPC1L1* rs3187907 NPY rs16475 *OLR1* rs11053646 *PDK*<sub>4</sub> rs12668651 *PPARGC1A* rs11734408 *PPARGC1A* rs2970865 PPARGC1B rs17572019 *PIK3R1* rs895304 *PLTP* rs378114

*ADRA2A* rs521674 *AGER* rs184003  $AGT \; \mathrm{AGT}$ *AKT2* rs7254617 *APOA4* rs5110 *APOA5* rs1729411 *APOB* rs1042031 *APOBEC1* rs7973596 *APOE* rs429358 APOL1 rs136175 *ASIP* rs819136 *C3* rs1047286 *CAPN10* rs2953171 *CARTPT* rs11575893 *CPT1A* rs4930610 CREBBP rs8046065 *ENPP1* rs1044558 *ESR1* rs1709183 *EXT2* rs3740878 *FABP2* rs10034661 *FABP4* rs1054135 *FOXO1* rs17446614 *FXN* rs2498429 *GAD2* rs2236418 *GAD2* rs928197 *GCK* rs1799884 *GIPR* rs1800437 *HK2* rs3755452 *HSD11B1* rs932335 *IDE* rs2249960 *IGF1* rs2195239 *IGF2* rs2230949 *INSR* rs1366234 *IRS2* rs1865434 *LCAT* rs4986970 LDLR rs2738465 *LEPR* rs3790419 *LIPC* rs8034802 *LMNA* rs577492 *LPIN1* rs2577261 *LPIN2* rs607549 LPL rs328 *LRP2* rs4667591 *LRP2* rs16856843 *MKKS* rs221666 *NEUROD1* rs8192556 *NOS3* rs11771443 *NPC1L1* rs217434 *NR1H3* rs12221497 *OLR1* rs3912640 *MTTP* rs982424 *PPARGC1A* rs2970869 *PPARGC1B* rs2010994 PPARGC1B rs26125 *PLA2G2D* rs10916711 *PON1* rs854551

*ADRA2A* rs1800544 *AGER* rs1035798 AGT rs4762*APOA1* rs5069 *APOA4* rs2239013 *APOA5* rs2266788 *APOB* rs676210 *APOC3* rs2070667 *APOH* rs3176975 *APOL1* rs136176 *ASIP* rs819162 *C3* rs2230199 *CAPN10* rs2953166 *CARTPT* rs10515114 *CETP* rs4783962 *CYP7A1* rs11786580 *ENPP1* rs7754859 *ESR1* rs1801132 *FABP1* rs2197076 *FABP2* rs10957056 *FABP6* rs2277954 *FOXO1* rs7986407 *GABRA6* rs7704209 *GAD2* rs7071922 *GCK* rs2268572 *GHRL* rs4684677 *GIPR* rs10423928 *HMGCR* rs3761738 *MARK2P9* rs2209972 *IDE* rs11187025 *IGF1* rs1019731 *INSIG1* rs9768687 *INSR* rs7254060 *KCNIP2* rs10883689 *LCAT* rs1109166 LEPR rs3790433 LEPR rs8179183 LIPC rs6078LMNA rs4641 *LPIN1* rs1050800 *LPIN2* rs650543 *LPL* rs10099160 *LRP2* rs2075252 *LRP5* rs312015 *MTTP* rs3792683 *NEUROD1* rs16867467 *NOS3* rs3918227 *NPC1L1* rs217420 *NR1H3* rs7120118 *PCK1* rs6070157 *PPARGC1A* rs3774923 *PPARGC1A* rs4383605 PPARGC1B rs741581 *PIK3R1* rs706713 *PLIN1* rs894160 PON2 rs7493

<i>PON2</i> rs10487133	<i>PON2</i> rs4729189	<i>PDZK1</i> rs1284300	<i>PPARA</i> rs12330015
PPARA rs4253728	PPARA rs1800206	PPARA rs6007662	PPARA rs4253778
PPARA rs6008259	<i>PPARD</i> rs2267665	<i>PPARD</i> rs2076169	PPARG rs6809631
PPARG rs10865710	PPARG rs1801282	PPARG rs3856806	<i>PRKAA1</i> rs17239241
<i>PRKAA1</i> rs3805492	<i>PRKAA1</i> rs10074991	<i>PRKAA1</i> rs466108	<i>PRKAA2</i> rs11206887
PRKAG3 rs6436094	PTPN1 rs6126033	PTPN1 rs6020608	PTPN1 rs2230604
FAM65C rs914458	PYY rs162430	PYY rs231461	<i>RETN</i> rs1862513
RETN rs3219177	<i>RETN</i> rs3745367	RXRB rs6531	RXRB rs2076310
RXRG rs17469611	SCARB1 rs701106	SCARB1 rs10846744	SCD rs3870747
SERPINE1 rs6090	<i>SERPINE1</i> rs2227657	SERPINE1 rs2227672	SERPINE1 rs2227692
<i>SLC25A20</i> rs4974088	<i>SLC27A4</i> rs7030121	<i>SLC27A4</i> rs6478827	SLC27A5 rs4801275
SLC2A2 rs5398	SLC2A2 rs5406	SLC2A2 rs10513684	<i>SLC2A2</i> rs5404
SLC2A2 rs5400	SLC2A4 rs5412	SLC2A4 rs5415	<i>SLC39A5</i> rs17118403
SLC39A6 rs1944319	<i>SLC6A14</i> rs12720074	SLC27A6 rs174006	<i>SOS1</i> rs2168043
<i>SOS1</i> rs7577088	SREBF1 rs11868035	<i>SREBF2</i> rs4501042	SREBF2 rs11702960
<i>SREBF2</i> rs4822063	SREBF2 rs5996080	SREBF2 rs4822062	STAT3 rs1053005
STAT3 rs2306580	STAT3 rs8069645	<i>HNF1A</i> rs3999413	TCF7L2 rs17685538
TCF7L2 rs1885510	TCF7L2 rs290481	TGFBI rs4572	TNF rs1799964
TNF rs1799724	TNF rs1800750	TNF rs1800629	<i>UCP1</i> rs11932232
<i>UCP1</i> rs7688743	<i>UCP1</i> rs1800592	<i>UCP2</i> rs17132534	UCP3 rs7930460
<i>UCP3</i> rs15763	UCP3 rs1800849	TSTD1 rs2073658	USF1 rs2774276
<i>USF1</i> rs2073655	<i>USF1</i> rs2073653	<i>USF1</i> rs1556259	VLDLR rs1454626
VLDLR rs1869592	VLDLR rs2290465	VLDLR rs8210	<i>ABCA1</i> rs2482432
<i>ABCA1</i> rs2065412	ABCA1 rs2230806	ABCA1 rs4743764	ABCA1 rs2740487
<i>ABCA1</i> rs1800977	<i>ABCA12</i> rs2225063	<i>ABCA12</i> rs4533467	ABCA12 rs4673937
<i>ABCA12</i> rs6744831	ABCA12 rs1523718	ABCA12 rs12464205	ABCC8 rs757110
ABCC8 rs916827	ABCC8 rs1799854	ABCC8 rs2040653	ABCC8 rs2237967
ABCC9 rs1283816	ABCC9 rs864360	ABCC9 rs829080	ABCC9 rs7301876
ABCC9 rs1517276	ABCC9 rs2138723	ABCD2 rs11172721	ABCG4 rs668033
ABCG5 rs2278356	ABCG5 rs10205816	ABCG5 rs4148187	ABCG5 rs4148182
ABCG5 rs4131229	ABCG5 rs13396273	ABCG5 rs3806471	ABCG5 rs3806470
ABCG8 rs6752551	ABCG8 rs4952689	ABCG8 rs4953028	ACACA rs1266182
ACACB rs6606697	ACACB rs2268387	ACACB rs3742026	ACACB rs2284685
ACACB rs3742023	<i>THYN1</i> rs570113	THYN1 rs1048761	ACAD8 rs473041
<i>ACADL</i> rs2286963	<i>ACADL</i> rs3764913	<i>ACADM</i> rs1251079	<i>ACADM</i> rs1250876
ACAT1 rs11212524	<i>ACAT1</i> rs10890819	<i>ACAT2</i> rs927450	ADIPOQ rs822395
ADIPOQ rs1063538	ACE rs4291	ACE rs4295	ACE rs4343
ACE rs4344	ACE rs4359	<i>ACE</i> rs4461142	ACE rs8066276
ACSL1 rs6552828	ACSL1 rs13120078	ACSL1 rs12503643	<i>ADD1</i> rs3775068
ADIPOR1 rs2275737	ADIPOR1 rs10920533	<i>ADIPOR2</i> rs6489323	<i>ADIPOR2</i> rs1058322
<i>ADRB2</i> rs1042713	<i>AGT</i> rs2067853	<i>AGT</i> rs2478523	AGT rs699
AGT rs5051	APOA1 rs5070	APOA2 rs5082	APOB rs693
<i>APOB</i> rs679899	<i>APOB</i> rs1367117	APOB rs512535	APOBEC1 rs9651863
APOE rs405509	APOE rs440446	APOH rs6933	APOL1 rs136147
APOLI rs3886200	<i>APOL6</i> rs9610329	<i>C3</i> rs11569562	$C_3$ rs344550
$C_3 rs 2241393$	$C_3 rs/25/062$	GPR35 rs2953161	<i>CARTPT</i> rs6453132
CD3b rs1984112	CD3b rs1/61667	CD3b rs3211816	CD3b rs3211931
CD30 rs1049673	OPIIA rs2305508	CFTD = 7005004	OEIP rs1/231506
CETP rs4/83961	CEIP rso20299	CEIP rs/205804 CEEPPP = 100000	OEIP rs4/84/44 OEEDDD = 0000000
$\cup EIF$ rs5882 CVD7/A1 ==28000007	OF12 rs1/99821	UREDDP rs129968	$\cup KEBBP$ rsv86528
$\cup I \Gamma / A I$ rs380860/	<i>UIP (AI</i> <b>rs</b> 3808607 FCD1 ma1514240	ENPPI rsobo344 EQD1 magnetize	ESRI TS532010 ESR1 mc6007070
ESRI ISZZ34093 $FSR1 ra19940E1$	ESRI IS1514340 EARD1 maggad 1992	ESRI IS055/1/1 $EARD1 ma2070001$	ESRI ISO92/0/2 EARD0 materials
$E_{DN1}$ 151004001 FARD0 referred 1	FADFI 152241003	FADF1 152510501 FARD6 ro10502547	FADF 2 181011020
FADF & 180001041 FA BD9 πα7750020	$\frac{FADF}{2} \frac{152252000}{FOYC2}$	$\frac{FADFU}{FVN} = 2200202$	FYN 77961007
$\frac{1}{CABBA6} = \frac{1}{CABBA6} = \frac{1}$	$\Gamma \cup A \cup Z$ $\Gamma \cup A \cup Z$ $C \land B D \land C \land T \to T$	$\frac{\Gamma A IV I S 23 U 33 33}{C C K \pi 23 C 6 8 E 7 4}$	1'AN 15/00199/
GADNAU ISODDDDDD	GADNAU IS3219151	GUN 182200014	GUN 182908290

GHRL rs35683	<i>P3H3</i> rs3759348	<i>P3H3</i> rs4963517	<i>GNB3</i> rs5440
GNB3 rs5443	<i>GYS1</i> rs2270938	<i>GYS1</i> rs5464	<i>HHEX</i> rs1111875
HK2 rs656489	<i>HK2</i> rs10496195	<i>HK2</i> rs4241264	<i>HMGCR</i> rs10038095
<i>HNF4A</i> rs3212183	<i>HNF4A</i> rs6017340	<i>HNF4A</i> rs6031596	<i>HNF4A</i> rs3818247
<i>LIPE</i> rs1206034	<i>IDE</i> rs2149632	<i>IDE</i> rs1544210	IGF1 rs6214
<i>IGF1</i> rs7136446	<i>IGF1R</i> rs4966014	<i>IGF1R</i> rs6598541	<i>IGF1R</i> rs2715428
IGF1R rs2229765	<i>IGF1R</i> rs7166565	<i>IGF1R</i> rs2593053	<i>IGF2</i> rs734351
<i>IGF2</i> rs3213216	<i>IGF2</i> rs3741211	<i>IKBKB</i> rs3747811	<i>IKBKB</i> rs5029748
IKBKB rs10958713	<i>IKBKB</i> rs3747811	<i>IKBKB</i> rs5029748	<i>IKBKB</i> rs10958713
IL6 rs1800797	<i>IL6</i> rs1800795	<i>IL6</i> rs2069832	<i>INSIG1</i> rs9769506
<i>INSIG1</i> rs9770068	<i>PDX1</i> rs4581569	<i>MTTP</i> rs10516445	<i>INSR</i> rs7248104
<i>INSR</i> rs17254521	<i>INSR</i> rs1346490	<i>INSR</i> rs919275	<i>PDX1</i> rs4581569
IRS2 rs2289046	IRS2 rs4771646	<i>KCNIP2</i> rs874885	<i>KCNJ11</i> rs2285676
<i>KCNJ11</i> rs5210	<i>KCNJ11</i> rs5215	<i>KCNJ11</i> rs5219	LDLR rs5930
LDLR rs688	<i>LEP</i> rs13228377	<i>LEP</i> rs11763517	<i>LEP</i> rs11761556
<i>LEPR</i> rs2025805	LEPR rs6673324	<i>LEPR</i> rs1137101	LEPR rs12067936
<i>LEPR</i> rs1805096	<i>LIPC</i> rs11629736	<i>LIPC</i> rs8028759	LIPC rs6083
LIPG rs2000813	LIPG rs2097055	LIPG rs2276269	LIPG rs6507931
<i>LIPH</i> rs6788865	<i>LIPH</i> rs9790230	<i>LPIN1</i> rs4315495	<i>LPIN1</i> rs7595221
<i>LPIN1</i> rs10209969	LPIN2 rs641287	<i>LPIN2</i> rs1785282	LPL rs1534649
LRP2 rs2161039	LRP2 rs831022	LRP2 rs700550	LRP2 rs2544377
LRP2 rs6716834	LRP2 rs4668136	LRP2 rs3755166	LRP5 rs4988300
LRP5 rs7944040	LRP5 rs638051	LTA rs915654	<i>LTA</i> rs2239704
LTA rs909253	LTA rs1041981	LTA rs915654	LTA rs2239704
<i>LTA</i> rs909253	<i>LTA</i> rs1041981	<i>MKKS</i> rs6077783	<i>SLX4IP</i> rs6039932
<i>MTTP</i> rs3816873	<i>MTTP</i> rs1057613	NCOA1 rs920418	NCOA1 rs2033861
<i>NEUROD1</i> rs1801262	NFATC3 rs7190307	<i>NOS3.1</i> NOS3.1	NOS3 rs1800783
<i>NPY</i> rs16129	<i>NR1H2</i> rs17373080	NR1H2 rs1405655	<i>OLR1</i> rs10505755
<i>OLR1</i> rs1050283	<i>TMEM52B</i> rs2742113	<i>PDK</i> / rs3807891	<i>MTTP</i> rs3811800
PPARGC1A rs8192678	PPARGC1A rs2970848	PPARGC1A rs4235308	PPARGC1A rs2946385
<i>PPARGC1A</i> rs2970870	PPARGC1B rs10060424	<i>PPARGC1B</i> rs10747516	PPARGC1B rs9324628
<i>PIK3R1</i> rs34303	<i>PIK3R1</i> rs10940160	<i>PLA2G2D</i> rs584367	<i>PLIN1</i> rs1052700
<i>PLTP</i> rs394643	<i>PON1</i> rs854549	<i>PON1</i> rs662	<i>PON1</i> rs854560
<i>PON1</i> rs2299261	<i>PON1</i> rs2299262	<i>PPARA</i> rs135539	PPARG rs2972164
PPARG rs709154	<i>PRKAA2</i> rs2746342	<i>PTPN1</i> rs6012953	<i>PTPN1</i> rs3787334
<i>PTPN1</i> rs2143511	<i>PTPN1</i> rs2038526	<i>PTPN1</i> rs2426159	PYYrs1058046
RETN rs3745369	<i>RXRA</i> rs1805352	<i>RXRA</i> rs4240705	<i>RXRA</i> rs1805343
RXRG rs283690	RXRG rs2134095	RXRG rs157865	<i>RXRG</i> rs2281985
SCAP rs12636851	SCAP rs6800271	<i>SCARB1</i> rs838891	<i>SCARB1</i> rs5888
SCARB1 rs989892	SCARB1 rs4765616	SCARB1 rs10846748	<i>SCARB1</i> rs10773109
SCARB1 rs3924313	SCARB1 rs10773111	SCARB1 rs11615630	SCD rs3071
SCD rs2234970	SCD rs3978768	<i>SERPINE1</i> rs2227631	<i>SLC25A14</i> rs2235800
SLC25A20 rs4974085	<i>SLC27A1</i> rs4808650	<i>SLC27A1</i> rs10423221	SLC27A1 rs2278280
SLC27A2 rs7176501	SLC27A2 rs2278167	<i>SLC27A2</i> rs1365508	SLC27A3 rs10158450
<i>SLC2A4</i> rs5435	SLC39A6 rs8092264	<i>SLC39A6</i> rs2236718	SLC39A6 rs948419
SLC6A14 rs7059155	SLC6A14 rs2071877	<i>SLC6A14</i> rs2011162	<i>SLC27A6</i> rs2577531
SLC27A6 rs185411	SLC27A6 rs1181965	SOS1 rs2888586	<i>SREBF1</i> rs9902941
TOM1L2 rs6502618	SREBF2 rs2267443	STAT3 rs744166	<i>HNF1A</i> rs1169288
<i>HNF1A</i> rs1169302	<i>HNF1A</i> rs2464196	<i>HNF1A</i> rs1169307	HNF1A rs735396
<i>TCF7L2</i> rs12255372	<i>TCF7L2</i> rs3814573	<i>TCF7L2</i> rs7903146	<i>TGFB1</i> rs4803455
TGFB1 rs2241715	<i>UCP1</i> rs12502572	<i>UCP2</i> rs660339	<i>UCP2</i> rs659366
<i>USF1</i> rs2516840	<i>VLDLR</i> rs2242103	<i>VLDLR</i> rs1454627	<i>VLDLR</i> rs2242104

Table 2: Complete list of the 712 categorical genotypic SNP variables analysed (and the associated genes). The first 371 SNP variables (by row) are binary variables and the remaining 341 SNP variables are nominal with 3 levels.

## 2 Mixture of Factor Analysers for Mixed Data, with Variable Selection – a Simulation Study

#### 2.1 Motivation

In order to assess the performance of the mixture of factor analysers for mixed data (MFA-MD) model with incorporated variable selection algorithm, and to assess the likelihood approximation and associated BIC-MCMC model selection criterion, a simulation study was conducted. Mixed type data were simulated from an MFA-MD model as detailed below. A range of MFA-MD models with varying values of G and Q was then fitted to the simulated data and the results analysed.

#### 2.2 Data Simulation

In order to simulate mixed type data, continuous data were first simulated from a mixture of G = 2 Gaussian distributions:

$$f(\underline{z}_i) = \sum_{g=1}^G \pi_g \text{MVN}_D\left(\underline{\mu}_g, \Lambda_g \Lambda_g^T + \Psi\right)$$

The matrices  $\Lambda_g$  were filled in with values simulated from a standard Gaussian distribution. The diagonal entries of  $\Psi$  corresponding to continuous variables were specified by taking the absolute value of Gaussian deviates with mean 1 and variance 1. The continuous data, which acts as the 'latent' continuous data of the MFA-MD model, were then transformed into 'observed' mixed type data using the criteria described in Section 3. A data set consisting of N = 500 observations on J = 738 variables was simulated. The variable types were divided as follows; 26 were continuous, 371 were binary and 341 were nominal. Thus the simulated data are similar in nature to the LIPGENE data set.

Two simulation settings were considered, both involving a mixture of two Gaussian distributions. The mixture components were well separated in the first simulation setting, but not well separated in the second. The degree of separation was achieved by specifying values for  $\underline{\mu}_1$  and setting  $\underline{\mu}_2$  equal to the negative of  $\underline{\mu}_1$ . Values for each entry of  $\underline{\mu}_1$  were simulated from a Gaussian distribution with mean 3 and variance 1 for the well separated case but with mean 1.5 and variance 1 for the less well separated case. The mixing weights  $\pi_g$  were set equal to 0.5 so that clusters were approximately equal in size. The dimension of the latent trait Q was specified to be 2.

In order to test the variable selection algorithm, 600 of the J = 738 variables were selected at random and the columns of the data corresponding to these variables were replaced with standard Gaussian deviates so that they did not contain clustering information. The variable selection algorithm should identify these noise variables and remove them from the model. Further the BIC-MCMC criterion, involving the approximated likelihood function, should select the generating G = Q = 2 model.

#### 2.3 Results

A range of MFA-MD models, with  $G = 1, \ldots, 4$  and  $Q = 1, \ldots, 3$ , were fitted to the simulated data sets. Each MCMC chain burned in for 5000 iterations, with a further

50,000 iterations, thinned every 50th. The thresholds for the variable selection algorithm  $\varepsilon$  were set at 0.9 for continuous variables and 0.95 for categorical variables. The BIC-MCMC criterion was evaluated for each model and the model with the largest BIC-MCMC selected as optimal. Further, the set of retained variables in the optimal model and the set of known clustering variables were compared.

#### 2.3.1 Simulation setting 1: Well Separated Mixture Components

Table 3 details the resulting BIC-MCMC values from fitting the MFA-MD models to the well separated simulated data. The true generating G = 2 and Q = 2 MFA-MD model has the largest BIC-MCMC value. The cluster memberships returned by the optimal MFA-MD model correspond exactly with the true component labels (Table 4). The variable selection algorithm correctly retained the 138 clustering variables, along with an additional 135 of the noise variables. Examination of the cluster means of the retained variables suggests that a less conservative threshold  $\varepsilon$  would have removed the retained non-clustering variables.

### 2.3.2 Simulation setting 2: Less Well Separated Mixture Components

In the case of the less well separated components, the similar behaviour was observed. Table 5 details the BIC-MCMC values, and the true generating G = Q = 2 model is selected as optimal. The resulting MFA-MD model clustered all the observations correctly, and retained 271 variables, of which 134 were non-clustering variables. One true clustering variable was also removed during the variable selection process.

The simulation study suggests that the approaches taken to variable selection and likelihood approximation, along with the BIC-MCMC model selection tool, perform well with the context of the MFA-MD model. Conservative values of the variable selection threshold  $\varepsilon$  are recommended to ensure all clustering variables are retained.

Table 3:	The	BIC-I	MCMC	C values	for	each	of	${\rm the}$	MFA	-MD	models	fitted	$\operatorname{to}$	the	well
separated	simu	ilated	data.	The opt	imal	l mod	el i	s sho	own ir	n bol	<u>d.</u>				

			Q	
		1	2	3
	1	-661361	-662344	-663317
G	2	-592824	-592371	-605088
	3	-599299	-599256	-600558
	4	-605993	-608576	-608571

Table 4: A cross tabulation of cluster membership versus the true cluster labels for the simulated data set with well separated clusters.

	True Component 1	True Component 2
Cluster 1	242	0
Cluster 2	0	258

Table 5: The BIC-MCMC values for each of the MFA-MD models fitted to the less well separated simulated data. The optimal model is shown in bold.

			Q	
		1	2	3
	1	-663809	-664818	-665927
G	2	-595978	-595751	-596256
	3	-598852	-600665	-602721
	4	-604278	-605234	-606663

# 3 Derivation of the Full Conditional Posterior Distributions

### 3.1 Mixing Weights

A Dirichlet prior distribution is used for the mixing weights  $\underline{\pi}$ .

 $\underline{\pi} \sim \text{Dirichlet}(\underline{\alpha})$ 

This is a conjugate prior which leads to a Dirichlet full conditional posterior.

$$h(\underline{\pi}|\dots) \propto \prod_{g=1}^{G} \pi_g^{(n_g + \alpha_g) - 1}$$
  
 $\Rightarrow \underline{\pi}|\dots \sim \text{Dirichlet}(\underline{\delta}_{\pi})$ 

where  $\underline{\delta}_{\pi} = (n_1 + \alpha_1, \dots, n_G + \alpha_G)$  and  $n_g = \sum_{i=1}^N \ell_{ig}$ .

### 3.2 Allocation Vectors

A priori the allocation vectors,  $\underline{\ell}_i$  are assumed to be Multinomial $(1, \underline{\pi})$  distributed. Thus the full conditional posterior is also multinomial.

$$h(\underline{\ell}_{i}|\ldots) \propto \prod_{g=1}^{G} \left\{ \pi_{g} \left[ \prod_{j=1}^{A} N(z_{ij}|\underline{\tilde{\lambda}}_{gj}^{T}\underline{\tilde{\theta}}_{i}, \psi_{jj}) \right] \left[ \prod_{j=A+1}^{B} \prod_{k=1}^{K_{j}} N(z_{ij}|\underline{\tilde{\lambda}}_{gj}^{T}\underline{\tilde{\theta}}_{i}, 1)^{\mathbb{I}\{z_{ij}<0|y_{ij}\}} \right] \times \left[ \prod_{j=A+B+1}^{J} \prod_{k=1}^{2} \prod_{s=0}^{2} N(z_{ij}^{k-1}|\underline{\tilde{\lambda}}_{gj}^{k-1^{T}}\underline{\tilde{\theta}}_{i}, 1)^{\mathbb{I}(y_{ij}=s)} \right] \right\}^{\ell_{ig}}$$
$$\Rightarrow \underline{\ell}_{i}|\ldots \sim \text{Multinomial}(\underline{p})$$

where

$$p_{g} = \left\{ \pi_{g} \left[ \prod_{j=1}^{A} N(z_{ij} | \underline{\tilde{\lambda}}_{gj}^{T} \underline{\tilde{\theta}}_{i}, \psi_{jj}) \right] \left[ \prod_{j=A+1}^{B} \prod_{k=1}^{K_{j}} N(z_{ij} | \underline{\tilde{\lambda}}_{gj}^{T} \underline{\tilde{\theta}}_{i}, 1)^{\mathbb{I}\{z_{ij} < 0 | y_{ij}\}} \right] \right. \\ \times \left[ \prod_{j=A+B+1}^{J} \prod_{k=1}^{2} \prod_{s=0}^{2} N(z_{ij}^{k-1} | \underline{\tilde{\lambda}}_{gj}^{k-1}^{T} \underline{\tilde{\theta}}_{i}, 1)^{\mathbb{I}(y_{ij}=s)} \right] \right\} \\ \times \left\{ \sum_{g=1}^{G} \pi_{g} \left[ \prod_{j=1}^{A} N(z_{ij} | \underline{\tilde{\lambda}}_{gj}^{T} \underline{\tilde{\theta}}_{i}, \psi_{jj}) \right] \left[ \prod_{j=A+1}^{B} \prod_{k=1}^{K_{j}} N(z_{ij} | \underline{\tilde{\lambda}}_{gj}^{T} \underline{\tilde{\theta}}_{i}, 1)^{\mathbb{I}\{z_{ij} < 0 | y_{ij}\}} \right] \\ \times \left[ \prod_{j=A+B+1}^{J} \prod_{k=1}^{2} \prod_{s=0}^{2} N(z_{ij}^{k-1} | \underline{\tilde{\lambda}}_{gj}^{k-1}^{T} \underline{\tilde{\theta}}_{i}, 1)^{\mathbb{I}(y_{ij}=s)} \right] \right\}^{-1}$$

### 3.3 Marginal Variance Parameters

An inverse gamma prior distributions is specified for the diagonal elements of  $\Psi$  which correspond to continuous items j.

$$\psi_{jj} \sim \mathcal{G}^{-1}\left(\beta_1, \beta_2\right)$$

This is a conjugate prior and leads to an inverse gamma posterior distribution.

$$\begin{split} h(\psi_d|\dots) &\propto \prod_{i=1}^N \left[ N\left( z_{id} | \underline{\lambda}_{gd}^T \tilde{\underline{\theta}}_i, \psi_d \right) \right]^{\ell_{ig}} \mathcal{G}^{-1} \left( \beta_1, \beta_2 \right) \\ &\propto \psi_d^{-\frac{N}{2}} \exp \left\{ -\frac{1}{2} \sum_{g=1}^G \sum_{i:\ell_{ig}=1} \psi_d^{-1} \left( z_{id} - \underline{\lambda}_{gd}^T \tilde{\underline{\theta}}_i \right)^2 \right\} \psi_d^{-\beta_1 - 1} \exp \left( -\frac{\beta_2}{\psi_d} \right) \\ &\propto \psi_d^{-\beta_1 - \frac{N}{2} - 1} \exp \left\{ -\frac{1}{2\psi_d} \sum_{g=1}^G \sum_{i:\ell_{ig}=1} \left( z_{id} - \underline{\lambda}_{gd}^T \tilde{\underline{\theta}}_i \right)^2 - \frac{\beta_2}{\psi_d} \right\} \\ &\propto \psi_d^{-\beta_1 - \frac{N}{2} - 1} \exp \left\{ \psi_d^{-1} \left[ \frac{1}{2} \sum_{g=1}^G \left( \underline{z}_{gd} - \tilde{\theta}_g \underline{\lambda}_{gd} \right)^T \left( \underline{z}_{gd} - \tilde{\theta}_g \underline{\lambda}_{gd} \right) + \beta_2 \right] \right\} \\ &\Rightarrow \psi_d | \dots \sim \mathcal{G}^{-1} \left\{ b_{1j}, b_{2j} \right\} \end{split}$$

where

$$b_{1j} = \beta_1 + \frac{N}{2}$$

and

$$b_{2j} = \left[\frac{1}{2}\sum_{g=1}^{G} \left(\underline{z}_{gd} - \tilde{\theta}_{g}\underline{\lambda}_{gd}\right)^{T} \left(\underline{z}_{gd} - \tilde{\theta}_{g}\underline{\lambda}_{gd}\right) + \beta_{2}\right]$$

### 3.4 Latent Traits

A priori the latent traits,  $\underline{\theta}_i,$  are assumed to be standard multivariate Gaussian distributed.

$$\underline{\theta}_i \sim \mathrm{MVN}_q \left( \underline{\theta}_i | \mathbf{0}, \mathbf{I} \right)$$

The resulting full conditional posterior is also multivariate Gaussian.

$$h(\underline{\theta}_{i}|\ldots) \propto \prod_{d=1}^{D} N\left(z_{id}|\underline{\tilde{\lambda}}_{gd}^{T}\underline{\tilde{\theta}}_{i},1\right) MVN_{q}\left(\underline{0},\mathbf{I}_{q}\right)$$

$$\propto \exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\left(z_{id}-\underline{\tilde{\lambda}}_{gd}^{T}\underline{\tilde{\theta}}_{i}\right)^{2}\right\} \exp\left\{-\frac{1}{2}\underline{\theta}_{i}^{T}\underline{\theta}_{i}\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\underline{\theta}_{i}^{T}\left[\Lambda_{g}^{T}\Lambda_{g}+\mathbf{I}_{q}\right]\underline{\theta}_{i}+\left[\left(\underline{z}_{i}-\underline{\mu}_{g}\right)^{T}\Lambda_{g}\right]\underline{\theta}_{i}\right\}$$

$$\Rightarrow \underline{\theta}_{i}|\ldots\sim MVN_{q}\left\{\underline{\mu}_{\theta},\Sigma_{\theta}\right\}$$

where

$$\underline{\mu_{\theta}} = \left[\Lambda_g^T \Lambda_g + \mathbf{I}_q\right]^{-1} \left[\Lambda_g^T \left(\underline{z}_i - \underline{\mu}_g\right)\right]$$

and

$$\Sigma_{\theta} = \left[\Lambda_g^T \Lambda_g + \mathbf{I}_q\right]^{-1}$$

### 3.5 Loadings Matrix and Mean

A multivariate Gaussian prior is specified for  $\underline{\tilde{\lambda}}_{gd}$ .

$$\underline{\tilde{\lambda}}_{gd} \sim MVN(q+1) \left(\underline{\tilde{\lambda}}_{gd} | \underline{\mu}_{\lambda}, \Sigma_{\lambda}\right)$$

This is a conjugate prior and so leads to a multivariate Gaussian full conditional posterior.

$$\begin{split} h(\tilde{\underline{\lambda}}_{gd}|\dots) &\propto \prod_{i=1}^{N} \left[ N\left(z_{id}|\underline{\tilde{\lambda}}_{gd}^{T}\underline{\tilde{\theta}}_{i},1\right) \right]^{\ell_{ig}} \left[ MVN_{(q+1)}\left(\underline{\tilde{\lambda}}_{gd}|\underline{\mu}_{\lambda},\Sigma_{\lambda}\right) \right] \\ &\propto \exp\left\{ -\frac{1}{2} \sum_{i:\ell_{ig}=1} \left( z_{id} - \underline{\tilde{\lambda}}_{gd}^{T}\underline{\tilde{\theta}}_{i} \right)^{2} \right\} \exp\left\{ -\frac{1}{2} \left( \underline{\tilde{\lambda}}_{gd} - \underline{\mu}_{\lambda} \right)^{T} \Sigma_{\lambda}^{-1} \left( \underline{\tilde{\lambda}}_{gd} - \underline{\mu}_{\lambda} \right) \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \underline{\tilde{\lambda}}_{gd}^{T} \left[ \Sigma_{\lambda}^{-1} + \underline{\tilde{\Theta}}_{g}^{T} \underline{\tilde{\Theta}}_{g} \right] \underline{\tilde{\lambda}}_{gd} + \underline{\tilde{\lambda}}_{gj}^{T} \left[ \underline{\tilde{\Theta}}_{g}^{T} \underline{z}_{gd} + \Sigma_{\lambda}^{-1} \underline{\mu}_{\lambda} \right] \right\} \\ &\Rightarrow \quad \underline{\tilde{\lambda}}_{gd} | \dots \sim MVN_{(q+1)} \left\{ \underline{\zeta}_{\lambda}, \Omega_{\lambda} \right\} \end{split}$$

where  $\underline{z}_{gd} = \{z_{id}\}$  for all respondents *i* in cluster *g*,  $\tilde{\Theta}_g$  is a matrix, the rows of which are  $\underline{\tilde{\theta}}_i$  for members of cluster *g* 

$$\underline{\zeta}_{\lambda} = \left[ \Sigma_{\lambda}^{-1} + \tilde{\Theta}_{g}^{T} \tilde{\Theta}_{g} \right]^{-1} \left[ \tilde{\Theta}_{g}^{T} \underline{z}_{gd} + \Sigma_{\lambda}^{-1} \underline{\mu}_{\lambda} \right]$$

and

$$\Omega_{\lambda} = \left[ \Sigma_{\lambda}^{-1} + \tilde{\Theta}_{g}^{T} \tilde{\Theta}_{g} \right]^{-1}.$$

## 4 Evolution of the variable selection phase



Figure 1: The evolution of the number of variables retained by the model over the duration of the variable selection phase of the model fitting algorithm.

# 5 Bayesian residual and Bayesian latent residual plots



Figure 2: Density estimates of the Bayesian residuals for the waist circumference variable for 50 randomly selected volunteers. The standard Gaussian density curve is shown by the black dashed line.



(b) Second latent dimension.

Figure 3: Density estimates of the Bayesian latent residuals for the rsrs2071877 SNP on the SLC6A14 gene, for 50 randomly selected volunteers. The standard Gaussian density curve is shown by the black dashed line.