Supplementary material for: Exome sequencing and the genetic basis of complex traits

Adam Kiezun[1,2,14], Kiran Garimella[2,14], Ron Do[2,3,14], Nathan O. Stitziel[4,2,14], Benjamin M. Neale[2,3,13], Paul J. McLaren[1,2], Namrata Gupta[2], Pamela Sklar[5], Patrick F. Sullivan[6], Jennifer L. Moran[2], Christina M. Hultman[7], Paul Lichtenstein[7], Patrik Magnusson[7], Thomas Lehner[8], Yin Yao Shugart[9], Alkes L. Price[2,10,11,15], Paul I.W. de Bakker[1,2,12,15], Shaun M. Purcell[13,15] & Shamil R. Sunyaev[1,2,15,16]

[1]*Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*

[2]*The Broad Institute of MIT and Harvard, Cambridge, MA, USA*

[3]*The Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA*

[4]*Division of Cardiovascular Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*

[5]*Department of Psychiatry, Friedman Brain Institute & Institute for Genomics and Multiscale Biology, Mount Sinai School of Medicine, New York, NY, USA*

[6]*Department of Genetics, University of North Carolina School of Medicine, Chapel Hill, NC, USA*

[7]*Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden*

[8]*Division of Neuroscience and Basic Behavioral Science, National Institute of Mental*

Health, Bethesda, MD, USA

[9]*Division of Intramural Research Program, National Institute of Mental Health, Bethesda, MD, USA*

[10]*Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA*

[11]*Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA*

[12]*Department of Medical Genetics and Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands*

[13]*Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA*

[14]*These authors contributed equally to this work.*

[15]*These authors jointly supervised this work.*

[16]*Corresponding author: ssunyaev@rics.bwh.harvard.edu*

Supplementary Note: Sweden National Cohort Collection - Schizophrenia: All participants (born in Sweden or another Nordic country) are recruited under protocols approved by the IRB at the Karolinska Institutet and with permission from the health board to which the potential subject was registered. All participants give written informed consent which does not prohibit sharing of genotypic data with other scientists. Cases are identified via the Swedish Hospital Discharge Register with discharge diagnoses of schizophrenia with two or more admissions. The sample is population-based covering all hospital-treated patients within three Swedish counties (Uppsala, Gästrikland and Västmanland).

Cases are interviewed about other lifetime medical conditions by the nurse who obtained consent and a blood sample. Controls are ascertained, with informed consent, frequency matched to cases by age, gender and county of residence. Controls are also identified from national population registers, and had never received a discharge diagnosis of schizophrenia or bipolar disorder. Prospective participants are sent an introductory letter explaining the study followed within two weeks by a telephone call to discuss the study and to gauge interest in participation. Interested subjects were sent an informational brochure and an appointment with a research nurse was be scheduled. At the appointment, informed consent was obtained and the research nurse then obtained a venous blood sample (2 $\times$ 10ml in EDTA tubes) using sterile technique. The tubes were transported to the Karolinska Institutet Biobank for DNA extraction. During a brief interview with the research nurse, cases are asked about the ancestries of their parents and grandparents, whether they have any history of a medical condition that could lead to phenocopies – a set of specific endocrine, autoimmune, and neurological disorders (e.g., epilepsy and brain damage) along with specific queries about licit and illicit drug use. Specific permission is obtained to re-contact the subject about participation in future studies, to contact their physicians, and to obtain their medical records. All data and samples were derived from previously collected materials, and no new human subjects were be enrolled in the research.

Inclusion/Exclusion criteria: For cases, individuals who have been hospitalized 2 or more times with schizophrenia discharge diagnosis on the Swedish Inpatient Hospital-

ization Register, alive, both parents born in Sweden, and provide informed consent are eligible. Individuals with hospital register diagnosis consistent with a medical or other psychiatric disorder that mitigates the schizophrenia diagnosis, or relationship closer then 2nd degree relative with another case are excluded. For controls, individuals matched to cases, never hospitalized with a discharge diagnosis of schizophrenia in the Inpatient Hospitalization register, alive, both parents born in Sweden, and provision of written informed consent are eligible. Control individuals with relation to any other case or control are excluded.

HIV controllers (n=163) were recruited through local outpatient clinics affiliated with the Ragon Institute of MGH, MIT and Harvard and collaborations with 335 health care providers and scientists across the US, Canada, Western Europe and Australia (`http://www.hivcontrollers.org/membermap`), as part of the International HIV Controllers Study. The respective institutional review boards (IRB) approved the study, and all subjects gave written informed consent. Persons were classified as HIV controllers if plasma HIV RNA was below 2000 copies/mL with a minimum of 3 determinations in the absence of antiretroviral therapy, spanning at least a 12-month period.

Chronically HIV infected non-controllers (n=21) were recruited from four trials of antiretroviral nave individuals performed through the AIDS Clinical Trials Group (ACTG protocols ACTG384, A5095, A5142 and A5202) within the United States. Eligible participants were adult patients with confirmed HIV-1 infection who had not taken prior an-

tiretroviral therapy. All individuals gave written informed consent and were consented under protocol A5128 for use of stored specimens for genetic testing.

Supplementary Table 1: SNP counts in 438 samples, stratified by allele count, filter status, and novelty (presence in dbSNP build 129). SNP filters have their greatest effect at low allele counts. The proportion of variants discarded in the allele count 1 or allele count 2 classes is greater than the proportion of variants discarded at all higher allele counts combined. The variants with low allele frequency are intuitively the ones most likely to be confused for machine noise and least likely to be discovered in large scale genotyping or sequencing projects like HapMap or 1000 Genomes. "Novel" SNPs are those absent from revision 129 of dbSNP. It is imperative that filters be carefully constructed so as to retain as much of the true low-frequency variation as possible without being too lax and flooding the dataset with false-positives.

Supplementary Table 2: Genotype counts, Ti/Tv, non-reference sensitivity (NRS), and non-reference discrepancy (NRD) rate for HIV, SCZ, and HapMap samples sequenced by Complete Genomics. As the metrics between the HIV and SCZ exomes are comparable to the expectations derived from the Complete Genomics data, the sensitivity and specificity of latter samples should provide insight into the quality of the former samples (given similar population backgrounds and subject to the variance incurred by hybrid capture). Here, NRS is high for all variants and around 50% for novel variants, indicating that most known variation is easily ascertained, but expected novels are more difficult to

recover. NRD is low for all and novel variants; for those variants that can be recovered, they are generally genotyped correctly. “Novel” SNPs are those absent from revision 129 of dbSNP.

By allele count	Filter	Novelty	Counts	# Ti	# Tv	Ti/Tv
1	unfiltered	all	39,602	27,051	12,551	2.16
		(novel)	(34,657)	(23,226)	(11,431)	(2.03)
	filtered	all	34,545	25,103	9,442	2.66
		(novel)	(29,694)	(21,333)	(8,361)	(2.55)
2	unfiltered	all	9,992	6,566	3,426	1.92
		(novel)	(7,237)	(4,473)	(2,764)	(1.62)
	filtered	all	7,661	5,711	1,950	2.93
		(novel)	(4,964)	(3,653)	(1,311)	(2.79)
> 2	unfiltered	all	57,262	39,536	17,726	2.23
		(novel)	(17,533)	(9,170)	(8,363)	(1.10)
	filtered	all	46,090	35,082	11,008	3.19
		(novel)	(7,846)	(5,692)	(2,154)	(2.64)

					Affymetrix 5.0/6.0 (HapMap sites only)			OMNI		
Sample Set	**# Samples**	**Novelty**	**# Variants**	**Ti/Tv**	**# Sites**	**% NRS**	**% NRD**	**# Sites**	**% NRS**	**% NRD**
HIV Elite Controllers	184	all (novel)	16,443 (1,283)	3.21 (2.72)	-	-	-	-	-	-
Schizophrenia controls	254	all (novel)	16,860 (1,272)	3.21 (2.92)	1,097 (0)	98.55 (0)	0.17 (0)	-	-	-
Complete Genomics HapMap samples	37	all (novel)	16,914 (1,269)	3.31 (2.86)	1,026 (0)	99.38 (0)	1.07 (0)	11,779 (570)	95.65 (57.72)	1.79 (1.12)