Some description of SNPs found by high throughput genome sequencing in the 1000 Genomes Project, and submitted to Illumina for the design of iSelect genotyping assays.

Tom Blackwell, University of Michigan, May 2, 2009

Summary:

The 150,000 SNPs we selected as candidates for Illumina genotyping assays fall into five general categories:

1. Functional: 55499 non-synonymous and splice site SNPs in Ensembl genes.
2. Contiguous: 18474 SNPs in four contiguous 1 Mb regions which cover reported GWAS signals for three or more disease traits. These four regions are in chromosome bands 9p21, 8q24, 6q23.3 and 1p13.2.
3. GWAS regions: 62477 SNPs in 100 smaller contiguous regions covering additional GWAS signals. These were chosen from the NHGRI table of published GWAS results compiled by Teri Manolio (www.genome.gov/26525384).
4. Isolated SNPs: 6940 isolated SNPs, comprised of either the single SNP reported for each of the remaining 954 GWAS signals (954 SNPs) or 5986 SNPs randomly chosen from all 1000 Genomes results, excluding those already selected in other classes.
5. Baylor Encode regions: 10482 SNPs in ten 100 kb regions that have been re-sequenced in 692 individuals from 10 populations worldwide for the HapMap 3 project by the Baylor College of Medicine HGSC. In these ten regions only, 1000 Genomes SNP calls have been augmented with 5758 SNPs called by Baylor from their own sequence data and with 4484 SNPs from dbSNP build 129.

After removing 1098 duplicates, including 665 non-synonymous and splice site SNPs which also appear in other categories, a non-redundant total of 152,774 candidate loci were submitted for assay design. Details of the selection for each category will be given below. All of the source data used in this compilation are publicly available on the web. Web references appear at the end of this summary.

Source of SNPs:

In December and February 2008-9, the 1000 Genomes project submitted a total of 11.5 M SNPs to the dbSNP public database. These were based on $>$ 21x coverage genome sequencing for two HapMap trio families (December) and on 2.5x coverage sequencing for a total of 105 unrelated individuals from four worldwide populations (February, these are also from HapMap). Please see .readme files at the 1000 Genomes ftp site for details of the methods and HapMap identifiers for the individuals sequenced. Provisional ss ('submitter SNP') numbers are now (May 1) available for all of these submissions; permanent rs numbers will be assigned in dbSNP build 132 later this year.

For the Illumina genotyping assays we excluded all X chromosome SNPs and all SNPs found to be tri- or tetra-allelic when calls were merged across populations. We added, at Richard Durbin's request, all SNP calls made for the YRI child (individual NA19240) by either center alone and excluded from the dbSNP submission.

The bulk of the data come from 9.8 M SNP calls based on low coverage sequencing. Additional data sources in the Baylor Encode regions are described in category 5 below. The exact numbers obtained from various sources are shown in Table 1.

Table 1: The pool of SNP calls available for inclusion in genotyping assays.

Key to column headings:

source	=	population subset used for 1000 Genomes SNP discovery
exclusive	=	number of 1000 Genomes autosomal SNP calls, exclusive of those in preceding categories
selected	=	number submitted for genotype assay design
inclusive	=	total autosomal SNP calls from this source, including SNPs which were already counted in preceding categories

source	exclusive	selected	inclusive
1000 Genomes SNP calls from low coverage sequence data for 105 individuals from three populations, February 2009 release:			
reported from all 3 popn's	2,002,589	21215	–
reported from 2 / 3 popn's	1,980,115	22319	–
CEU only	1,969,492	28302	5,619,647
YRI only	2,781,078	34522	5,850,263
CHB+JPT	1,082,797	15622	4,331,454
total low coverage	9,816,071	121,980	
1000 Genomes SNP calls from high coverage trio data, December 2008 releases:			
CEU trio	762,761	8399	3,852,959
YRI child, submitted	731,257	8503	3,464,673
YRI child, Sanger only	400,716	5028	768,685
YRI child, AB SOLiD only	237,516	2323	266,187
1000 Genomes grand total	11,948,321	146,233	
SNP calls from Baylor:	–	4505	5,758
SNP calls from dbSNP build 129:	–	2036	11,192,764
Submitted grand total:		152,774	

Notes: (1) The total numbers of SNPs in this table will differ from totals for the original releases due to excluding X chromosome and tri-allelic loci here. (2) It is possible that loci which are tri-allelic across populations, and hence excluded from the low coverage data, were restored if the same locus was called using either the CEU or YRI high coverage data. (3) 33 duplicate loci were accidentally submitted in the Baylor Encode regions. These have been removed from the counts shown here, but retained in the master submission file. Thus, there will be 152,807 rows in the master submission file rather than 152,774.

Target regions:

As we began the SNP selection process, I estimated that the 1000 Genomes Project has reported so far an average of roughly 4,200 SNPs per Mb of genomic sequence. A more precise number is 4,456 SNPs per Mb, if one uses a denominator of 2,681.3 Mb for the total length of autosomal sequence excluding assembly gaps

in the NCBI build 36 reference genome assembly. (This quantity is taken from the 'stats' page of the UCSC genome browser.) Using either number, and given Illumina's target of 100,000 SNPs, we would need to select regions covering a total length of 22 – 30 Mb of genomic sequence.

Five general categories:

(1) Functional: Initial interest focused on including all non-synonymous coding SNPs. In view of the complexities of plus and minus strands, intron phase and three possible reading frames, I was not confident that I could correctly determine which SNPs would produce synonymous or non-synonyomous changes in the coding sequence, on the short time scale required. However, we could not submit all 221,966 SNPs from the 1000 Genomes project that fall in or within 5 bp of Ensembl annotated exons (85 Mb). This would already be more SNPs than the manufacturing capacity available. At this point, the project was at an impasse.

The whole project would not have gone forward had not Jim Stalker at Sanger recognized that distinguishing synonymous from non-synonymous SNPs would be, in his words, "a pretty trivial use of the Ensembl Variation API." "You make a new Bio::EnsEMBL::Variation::AlleleFeature with your snp location and allele string, find any transcripts that overlap the snp, and then use a utility method called get_all_ConsequenceType from Bio::EnsEMBL::Utils::TranscriptAlleles to return the consequence of the AlleleFeature on that transcript. A script that wraps the whole thing is only a hundred lines or so."

Jim spent his Friday evening doing the analysis for us and tabulating the results. This classified the 221,966 exonic SNPs into 11 Ensembl functional categories as follows.

Table 2: Functional classification of 221,966 exonic SNPs.

NON_SYNONYMOUS_CODING	48307
STOP_GAINED	1084
STOP_LOST	73
ESSENTIAL_SPLICE_SITE	879
SPLICE_SITE	5156
SYNONYMOUS_CODING	38717
3PRIME_UTR	87330
5PRIME_UTR	17428
NO_CODING_TRANSCRIPTS	15483
INTRONIC	6354
INTERGENIC	1155

In the Ensembl help system, 'ESSENTIAL_SPLICE_SITE' is defined as 'the first 2 or last 2 basepairs of an intron', while 'SPLICE_SITE' means '1-3 bps into an exon or 3-8 bps into an intron'. In addition to protein coding genes, the Ensembl gene and exon annotations include a variety of specific types of functional RNA transcripts and pseudogenes. This explains the noncoding and intergenic classes above. In the end, we submitted all 55499 SNPs from the first five categories, of which 665 also satisfy other categories.

I am surprised by the ratio of non-synonymous to synonymous polymorphisms shown in this tally. Perhaps the surprise occurs because one usually sees this ratio in comparisons between species, where the total number of differences is much larger, and where both drift and selection have acted over much longer time scales.

(2) Contiguous: Four chromosome regions are strongly implicated in GWAS studies for three or more different phenotypes. These regions are: 9p21 (type 2 diabetes, myocardial infarction, coronary disease, intracranial aneurysm), 8q24 (multiple cancers), 6q23.3 (psoriasis, systemic lupus erythematosus, rheumatoid arthritis) and 1p13.2 (Crohn's disease, rheumatoid arthritis, type 1 diabetes).

For each region, a 1 Mb contiguous segment was chosen which covers the reported GWAS signals. Region boundaries were chosen with regard to the neighboring annotated genes. All 1000 Genomes SNPs within each region were submitted for genotyping assays. The region boundaries are shown below, along with the number of SNPs and the number of GWAS signals. It is hoped that these four regions may be useful for population genetics studies of linkage disequilibrium as well as for validating our methods for sequence based SNP discovery. In addition to these four regions, at 966 kb, GWAS region # 98 from category 3 (following) is almost as large as any of the 1 Mb regions.

Column 'total' in the table below gives the total number of 1000 Genomes SNPs in each region. Column 'else' shows the number of SNPs which also satisfy other categories. Column 'gw' gives the number of distinct SNPs (rs numbers) from the region listed in the NHGRI GWAS table; column 'addl' is the number of additional reports in the GWAS table for these SNPs, after the first for each SNP.

Table 3: Boundaries and contents of four 1 Mb contiguous chromosome regions.

band	chr	start	stop	total	else	gw	addl
9p21	9	21.75 Mb	22.75 Mb	4568	7	7	4
8q24	8	127.95 Mb	128.95 Mb	5495	5	10	5
6q23	6	137.60 Mb	138.60 Mb	4890	9	4	1
1p13.2	1	113.75 Mb	114.75 Mb	3521	33	2	7

(3) GWAS regions: Smaller regions were chosen surrounding an additional 150 GWAS signals, detailed in Table 4. Novel SNPs in strong linkage disequilibrium with an established association signal are of special interest, since any one of them might be a causative variant. With this in mind, the region boundaries were chosen to include all HapMap SNPs showing $r^2 \geq 50\%$ with a reported GWAS signal. Each region was then extended in both directions to the nearest HapMap SNP at least 0.02 cM beyond the boundary. This will include a neighboring recombination hotspot, if there is one nearby. Overlapping and adjacent regions were merged, reducing the total to 100 disjoint regions covering 15.77 Mb. There is wide variation in the region widths. The median width is 107 kb; their mean is 157.7 kb; but 1/3 of the regions are shorter than 75 kb and 1/3 are longer than 155 kb.

The 150 GWAS signals were selected favoring regions which show association with more than one phenotype or more than one study, while requiring a reported p-value of 5×10^{-8} or better. A deliberate effort was made to include a variety of phenotypes and a variety of authors, even at the cost of ignoring some results with stronger p-values. Associations at the HLA locus on chromosome 6p21.3 were explicitly avoided, yet one such region has appeared nonetheless. When merging adjacent regions, the result with the strongest p-value was chosen to represent each region. This masks some of the diversity sought among the phenotypes. The columns 'else', 'total', 'gw' and 'addl' in Table 4 have the same interpretation as for the four 1 Mb regions.

Table 4: Boundaries and contents of 100 GWAS regions.

row	chr	band	else	total	width	start	snp.pos	stop	rs	p-value	gw	addl	genes	author	journal	trait
1	1	1p36.13	3	529	123110	17578947	17594950	17702057	rs7538876	4e-12	1	0	PADI4,PADI6,RCC2,ARHGEF10L	Stacey	Nat Genet	Basal cell carcinoma (cutaneous)
2	1	1p36.12	0	251	60729	22550290	22571034	22611019	rs7524102	1e-16	2	2	Intergenic	Styrkarsdottir	Nat Genet	Bone mineral density (hip)
3	1	1p31.3	9	873	309505	62672382	62704220	62981887	rs1167998	2e-12	4	2	DOCK7	Aulchenko	Nat Genet	Triglycerides
4	1	1p31.3	3	638	117445	65826144	65878532	65943589	rs1892534	7e-21	2	0	LEPR	Ridker	Am J Hum Genet	C-reactive protein
5	1	1p13.3	5	62	17532	109612504	109623689	109630036	rs599839	1e-33	3	7	CELSR2,PSRC1	Sandhu	Lancet	LDL cholesterol
6	1	1q21.3	19	812	171810	150803779	150816642	150975589	rs4085613	7e-30	1	0	LCE3D,LCE3A	Zhang	Nat Genet	Psoriasis
7	1	1q23.3	0	582	111900	160286843	160352309	160398743	rs10494366	1e-10	1	0	NOS1AP	Arking	Nat Genet	QT interval prolongation
8	1	1q32.1	2	111	22805	201413209	201422505	201436014	rs4950928	1e-13	1	0	CHI3L1	Ober	N Engl J Med	YKL-40 levels
9	1	1q42.13	0	590	103067	227006918	227064458	227109985	rs801114	6e-12	1	0	RHOU	Stacey	Nat Genet	Basal cell carcinoma (cutaneous)
10	2	2p25.3	0	516	57563	587557	634953	645120	rs7561317	2e-18	2	1	TMEM18	Thorleifsson	Nat Genet	Weight
11	2	2p15.1	0	138	39546	60543252	60571547	60582798	rs1427407	6e-31	1	0	BCL11A	Menzel	Nat Genet	F-cell distribution
12	2	2p15	0	625	180000	61720000	61844742	61900000	rs1186868	7e-35	1	0	BCL11A	Uda	PNAS	Fetal hemoglobin levels
13	2	2p14	0	230	66207	66579957	66634957	66646164	rs2300478	3e-28	1	0	MEIS1	Winkelmann	Nat Genet	Restless legs syndrome
14	2	2q12.1	13	1008	182295	102279524	102324148	102461819	rs1420101	5e-14	2	0	IL1RL1	Gudbjartsson	Nat Genet	Plasma eosinophil count
15	2	2q24.3	4	339	64746	169460436	169471394	169525182	rs560887	1e-57	2	2	G6PC2	Prokopenko	Nat Genet	Fasting plasma glucose
16	2	2q32.3	1	170	84750	191597827	191672878	191682577	rs7574865	9e-14	2	0	STAT4	Hom	N Engl J Med	Systemic lupus erythematosus
17	2	2q35	0	495	73809	217572846	217614077	217646655	rs13387042	1e-13	1	0	Intergenic	Stacey	Nat Genet	Breast cancer
18	2	2q37.1	2	433	99500	233805923	233845149	233905423	rs3828309	2e-32	3	0	ATG16L1	Barrett	Nat Genet	Crohn's disease
19	3	3p14.3	0	247	51997	56822985	56840816	56874982	rs12485738	4e-27	1	0	ARHGEF3	Meisinger	Am J Hum Genet	Mean platelet volume
20	3	3q21.3	3	484	106583	129679190	129743240	129785773	rs4857855	9e-17	1	0	GATA2	Gudbjartsson	Nat Genet	Plasma eosinophil count
21	3	3q22.3	7	854	286954	139385394	139604812	139672348	rs9818870	7e-13	1	0	MRAS	Erdmann	Nat Genet	Coronary artery disease
22	3	3q23	10	1033	305226	142513233	142585523	142818459	rs6763931	1e-27	3	1	ZBTB38	Gudbjartsson	Nat Genet	Height
23	3	3q27.2	0	311	103279	186929732	186994381	187033011	rs4402960	9e-16	1	3	IGF2BP2	Saxena	Science	Type 2 diabetes
24	4	4p16.3	20	766	151906	1017307	1085281	1169213	rs3796619	3e-24	2	0	RNF212,SPON2	Kong	Science	Recombination rate (males)
25	4	4p16.1	13	3901	504259	9521199	9531265	10025458	rs16890979	7e-168	5	2	SLC2A9	Dehghan	Lancet	Serum urate
26	4	4q22.1	6	310	102359	89182740	89271347	89285099	rs2231142	3e-60	1	0	ABCG2	Dehghan	Lancet	Serum urate
27	4	4q25	0	568	129409	111824176	111929618	111953585	rs2200733	3e-41	2	1	PITX2,ENPEP	Gudbjartsson	Nature	Atrial fibrillation/atrial flutter
28	4	4q27	25	1942	576779	123201764	123728871	123778543	rs6822844	1e-14	3	1	KIAA1109,TENR,IL2,IL21	van Heel	Nat Genet	Celiac disease
29	5	5q11.2	446	1279	276247	56024535	56067641	56300782	rs889312	7e-20	1	0	MAP3K1	Easton	Nature	Breast cancer
30	5	5q33.1	0	121	20457	150442831	150458511	150463288	rs17728338	1e-20	1	0	TNIP1	Nair	Nat Genet	Psoriasis
31	5	5q33.3	8	770	249395	158509493	158650367	158758888	rs2082412	2e-28	4	0	IL12B	Nair	Nat Genet	Psoriasis
32	6	6p25.3	0	485	77309	335105	341321	412414	rs12203592	7e-127	3	1	IRF4	Han	PLoS Genet	Black vs. blond hair color
33	6	6p22.3	0	525	111213	20728290	20836710	20839503	rs6908425	9e-10	6	2	CDKAL1	Barrett	Nat Genet	Crohn's disease
34	6	6p22.3	0	261	52655	22206026	22247983	22258681	rs6939340	9e-15	1	0	FLJ22536,FLJ44180	Maris	N Engl J Med	Neuroblastoma
35	6	6q22.1	6	662	154860	26249354	26341366	26404214	rs10946808	4e-17	1	1	HIST1H1D	Lettre	Nat Genet	Height
36	6	6p21.33	3	123	11000	31534000	31539759	31545000	rs2395029	2e-26	1	1	HLA-C	Liu	PLoS Genet	Psoriasis
37	6	6p21.33	42	352	107000	31678000	31728499	31785000	rs3117582	5e-10	2	0	BAT3,MSH5	Wang	Nat Genet	Lung cancer
38	6	6p21.2	0	1067	260361	38399891	38473819	38660252	rs9296249	4e-18	2	0	BTBD9	Winkelmann	Nat Genet	Restless legs syndrome
39	6	6q23.3	3	852	218959	135290739	135460711	135509698	rs9399137	3e-36	1	0	Intergenic	Menzel	Nat Genet	F-cell distribution
40	6	6q25.1	4	288	60190	151971942	151990059	152032132	rs2046210	2e-15	3	2	C6orf97	Zheng	Nat Genet	Breast cancer
41	6	6q25.1	0	374	94321	152049654	152110057	152143975	rs1999805	2e-08	1	0	ESR1,C6orf97	Styrkarsdottir	N Engl J Med	Bone mineral density (spine)
42	7	7p15.1	0	558	124500	28102337	28147081	28226837	rs864745	5e-14	2	0	JAZF1	Zeggini	Nat Genet	Type 2 diabetes
43	7	7p13	2	241	50619	44188106	44202193	44238725	rs4607517	1e-25	1	0	GCK	Prokopenko	Nat Genet	Fasting plasma glucose
44	7	7q21.2	2	359	144981	92066336	92102346	92211317	rs2282978	8e-23	3	1	CDK6	Weedon	Nat Genet	Height
45	7	7q21.3	20	1463	386222	97494613	97654263	97880835	rs6465657	1e-09	1	0	LMTK2	Eeles	Nat Genet	Prostate cancer
46	7	7q22.3	0	279	72449	106092115	106159455	106164564	rs342293	1e-24	1	0	Intergenic	Soranzo	Blood	Mean platelet volume
47	7	7q32.1	4	865	208269	128347997	128505142	128556266	rs12537284	4e-19	2	0	IRF5,TNPO3	Harley	Nat Genet	Systemic lupus erythematosus
48	8	8q11.23	1	379	96063	55446657	55489644	55542720	rs10958409	1e-10	1	0	SOX17	Bilguvar	Nat Genet	Intracranial aneurysm
49	8	8q11.23	0	505	106557	55567255	55600077	55673812	rs9298506	2e-09	1	0	SOX17	Bilguvar	Nat Genet	Intracranial aneurysm
50	9	9q22.33	3	814	217748	99546029	99595930	99763777	rs965513	2e-27	1	0	FOXE1	Gudmundsson	Nat Genet	Thyroid cancer
51	9	9q34	26	1112	340076	122673769	122730060	123013845	rs3761847	4e-14	1	0	TRAF1-C5	Plenge	N Engl J Med	Rheumatoid arthritis
52	10	10p14	0	399	80530	8726910	8741225	8807440	rs10795668	3e-13	1	0	Intergenic	Tomlinson	Nat Genet	Colorectal cancer
53	10	10q21.1	10	632	160960	42926862	42932615	43087822	rs2742234	4e-18	1	0	RET,GALNACT-2,RASGEF1A	Garcia-Barcelo	PNAS	Hirschsprung's disease
54	10	10q11.23	1	374	95455	51136350	51219502	51231805	rs10993994	9e-29	1	1	MSMB	Eeles	Nat Genet	Prostate cancer
55	10	10q21.2	1	1668	330755	61697612	61849818	62028367	rs10994336	9e-09	1	0	ANK3	Ferreira	Nat Genet	Bipolar disorder
56	10	10q23.33	24	2105	636403	96208128	96697192	96844531	rs4086116	6e-12	2	0	CYP2C9	Cooper	Blood	Warfarin maintenance dose
57	10	10q24.2	1	264	56765	101262195	101281583	101318960	rs11190140	3e-16	2	1	NKX2-3	Barrett	Nat Genet	Crohn's disease
58	10	10q25.2	0	197	74381	114735800	114744078	114810181	rs7901695	1e-48	3	7	TCF7L2	Zeggini	Science	Type 2 diabetes
59	10	10q26.13	1	166	33146	123319419	123342307	123352565	rs2981582	2e-76	2	0	FGFR2	Easton	Nature	Breast cancer
60	10	10q26	0	141	28000	124200000	124210534	124228000	rs11200638	8e-12	1	0	HTRA1	DeWan	Science	Age-related macular degeneration
61	11	11p15.5	10	212	60006	1845531	1865582	1905537	rs3817198	3e-09	1	0	LSP1	Easton	Nature	Breast cancer
62	11	11q13.2	0	243	61712	68721150	68751073	68782862	rs7931342	2e-12	2	0	Intergenic	Eeles	Nat Genet	Prostate cancer
63	11	11q23.1	3	329	70118	110617544	110676919	110687662	rs3802842	6e-10	1	0	Intergenic	Tenesa	Nat Genet	Colorectal cancer
64	11	11q24.1	0	248	47991	122850730	122866607	122898721	rs735665	4e-12	1	0	GRAMD1B	Di Bernardo	Nat Genet	Chronic lymphocytic leukemia
65	12	12q13.2	4	415	168673	54640539	54768447	54809212	rs2292239	2e-20	4	1	ERBB3	Todd	Nat Genet	Type 1 diabetes
66	12	12q14.3	4	282	83527	64592708	64644614	64676235	rs1042725	3e-20	2	2	HMGA2	Lettre	Nat Genet	Height
67	12	12q24.31	14	833	184551	119789209	119909244	119973760	rs7310409	7e-17	5	1	HNF1A	Ridker	Am J Hum Genet	C-reactive protein
68	12	12q24.31	18	649	189456	120765958	120849966	120955414	rs7961894	7e-48	1	0	WDR66	Meisinger	Am J Hum Genet	Mean platelet volume
69	13	13q34	14	269	60000	112788000	112808035	112848000	rs561241	5e-16	1	0	MCF2L,F7,F10,PROZ	Yang	BMC Med Genet	Factor VII
70	14	14q22.2	1	220	52398	53450742	53480669	53503140	rs4444235	8e-10	1	0	BMP4	COGENT Study	Nat Genet	Colorectal cancer
71	15	15q13.1	8	625	226459	26001430	26203777	26227889	rs1667394	1e-241	4	4	OCA2	Sulem	Nat Genet	Blue vs brown eyes
72	15	15q13.1	0	36	37262	26772075	26805134	26809337	rs8033165	2e-12	1	1	Intergenic	Han	PLoS Genet	Black vs. red hair color
73	15	15q23	1	368	87390	67749791	67806044	67837181	rs7176508	5e-12	2	0	Intergenic	Di Bernardo	Nat Genet	Chronic lymphocytic leukemia
74	15	15q24.1	4	186	38029	71992921	72006635	72030950	rs3825942	3e-21	1	0	LOXL1	Thorleifsson	Science	Exfoliation glaucoma
75	15	15q24.3	1	324	70210	75710127	75750942	75780337	rs9652490	1e-09	1	0	LINGO1	Stefansson	Nat Genet	Essential tremor
76	15	15q25.1	13	1205	288170	76498437	76593078	76786607	rs8034191	5e-20	3	3	CHRNA3,CHRNA5,CHRNB4,PSMA4	Hung	Nature	Lung cancer
77	15	15q25.1	9	677	126677	76919385	77022501	77046062	rs3825932	3e-15	1	0	CTSH	Cooper	Nat Genet	Type 1 diabetes
78	16	16p13.13	1	1029	241830	10938243	11087374	11180073	rs12708716	3e-18	3	2	KIAA0350	Todd	Nat Genet	Type 1 diabetes
79	16	16p11.2	45	1438	555134	30731722	30955580	31286856	rs10871454	5e-34	4	0	VKORC1	Cooper	Blood	Warfarin maintenance dose
80	16	16q12.1	6	453	123356	49261663	49314382	49385019	rs2076756	1e-21	4	4	CARD15	Franke	PLoS ONE	Inflammatory bowel syndrome
81	16	16q12.1	0	565	120429	51091668	51143842	51212097	rs3803662	1e-36	1	1	TNCR9,LOC643714	Easton	Nature	Breast cancer
82	16	16q13	2	160	25959	55540050	55562980	55566009	rs1532624	9e-94	6	4	CETP	Aulchenko	Nat Genet	HDL cholesterol
83	17	17q11.2	1	561	226237	24687467	24727475	24913704	rs2138852	7e-28	1	0	TAOK1	Meisinger	Am J Hum Genet	Mean platelet volume
84	17	17q12	0	69	18314	33167554	33172153	33185868	rs4430796	1e-11	2	1	TCF2	Gudmundsson	Nat Genet	Prostate cancer
85	17	17q21	21	982	294608	35081598	35323475	35376206	rs7216389	9e-11	2	0	Intergenic	Moffatt	Nature	Asthma
86	17	17q24.3	2	834	143285	66612602	66620348	66755887	rs1859962	3e-10	1	1	Intergenic	Gudmundsson	Nat Genet	Prostate cancer
87	18	18p11.21	1	712	150804	12728413	12769947	12879217	rs2542151	5e-17	1	4	PTPN2	Barrett	Nat Genet	Crohn's disease
88	18	18q21.1	2	57	10256	44700838	44707461	44711094	rs4939827	8e-28	1	2	SMAD7	Tenesa	Nat Genet	Colorectal cancer
89	18	18q21.32	1	946	187624	55879013	56035730	56066637	rs12970134	5e-13	2	4	MC4R	Thorleifsson	Nat Genet	Weight
90	19	19q13.11	9	457	122997	38203614	38224140	38326611	rs10411210	5e-09	1	0	RHPN2	COGENT Study	Nat Genet	Colorectal cancer
91	19	19q13.32	4	331	93000	50047000	50114786	50140000	rs4420638	1e-60	6	12	APOE,APOC1,APOC4,APOC2	Kathiresan	Nat Genet	LDL cholesterol
92	19	19q13.3	8	306	71213	53881227	53898229	53952440	rs492602	5e-17	1	0	FUT2	Hazra	Nat Genet	Plasma level of vitamin B12
93	19	19q13.33	7	218	26491	56045214	56056435	56071705	rs2735839	2e-18	1	0	KLK3	Eeles	Nat Genet	Prostate cancer
94	20	20p13	1	265	48000	4602000	4628251	4650000	rs1799990	2e-21	1	0	PRNP	Mead	Lancet Neurol	Creutzfeldt-Jakob disease
95	20	20p12.3	1	618	111060	6244266	6352281	6355326	rs961253	2e-10	1	0	Intergenic	COGENT Study	Nat Genet	Colorectal cancer
96	20	20p11.22	0	1798	468312	21702301	21998503	22170613	rs1160312	1e-14	2	0	PAX1	Richards	Nat Genet	Male-pattern baldness
97	20	20q11.22	2	1096	382977	31948190	32202273	32331167	rs1015362	6e-37	2	2	ASIP	Sulem	Nat Genet	Burning and freckling
98	20	20q11.22	28	2672	966427	32368297	32635433	33334724	rs910873	1e-15	1	0	CDC91L1	Brown	Nat Genet	Melanoma
99	20	20q11.22	31	929	299583	33349306	33377622	33648889	rs6060373	2e-17	3	1	GDF5	Weedon	Nat Genet	Height
100	21	21q22.11	0	362	88603	34515109	34520998	34603712	rs9982601	6e-11	1	0	SLC5A3,MRPS6,KCNE2	Kathiresan	Nat Genet	Myocardial infarction (early onset)

(4) Isolated SNPs: For the remaining 954 rs numbers in the GWAS table, only the reported SNP itself was submitted for genotyping assays. These include 675 rs numbers whose best p-value is worse than 5×10^{-8}. 933 / 954 signals coincide with a SNP from the 1000 Genomes project. The remaining 21 were submitted using information from dbSNP version 129.

Also in this category of isolated SNPs are 5986 SNPs chosen completely at random from the 1000 Genomes collection, but excluding any SNPs submitted in other categories. The random selection used the R function 'sample()' following 'set.seed(82675)'. This sample is interpreted relative to the 1000 Genomes non-redundant ordering. [As a technical detail, the selection of these random SNPs excludes a fixed width window of 180 kb around most of the 100 GWAS regions, rather than the smaller windows subsequently determined.]

(5) Baylor Encode regions: The Baylor College of Medicine Human Genome Sequencing center has re-sequenced ten 100 kb regions in each of 692 individuals for the HapMap 3 and Encode III projects. This is a much deeper sampling from the human population than in the pilot phase of the 1000 Genomes project. Their procedure uses PCR amplification and conventional Sanger dideoxy sequencing.

The Baylor Encode region identifiers and region boundaries are shown below. We submitted a non redundant set of 10,482 SNPs which contains all 3962 SNPs from the 1000 Genomes project in these regions, all 5758 SNPs called by Baylor from their own sequencing data, and 4484 SNPs from dbSNP build 129 which have 'class' = 'single' and 'locType' either 'exact' or 'between'. 1253 SNPs coincide between the Baylor and 1000 Genomes calls, while 2469 SNPs coincide between dbSNP and either sequence based set, leaving a non redundant total of 10,482 SNPs in the Baylor Encode regions. Fu Li Yu from Baylor notes that five of the regions contain highly repetitive sequence in which it will be difficult to design unique SNP assays. Table 5 shows the marginal totals for each region and includes some double counting. Table 6, following, gives details of the overlaps among all three data sources.

Table 5: Boundaries and total SNP counts in the ten Baylor Encode regions.

region	chr	region start	region stop	source	Baylor	dbSNP	1000 G	else
ENm010	7	27,124,046	27,224,045	ENCODE I	1041	540	390	7
ENr321	8	119,082,221	119,182,220	ENCODE I	1098	601	468	0
ENr232	9	130,925,123	131,025,122	ENCODE I	840	656	439	3
ENr123	12	38,826,477	38,926,476	ENCODE I	748	631	549	2
ENr213	18	23,919,232	24,019,231	ENCODE I	899	561	349	1
ENr331	2	220,185,590	220,285,589	ENCODE III	0	341	381	3
ENr221	5	56,071,007	56,171,006	ENCODE III	567	310	433	0
ENr233	15	41,720,089	41,820,088	ENCODE III	28	491	78	1
ENr313	16	61,033,950	61,133,949	ENCODE III	0	325	430	0
ENr133	21	39,444,467	39,544,466	ENCODE III	460	388	445	11
				totals:	5758	4844	3962	28

Table 6: Non redundant counts of SNPs reported by each data source or combination of data sources within the ten Baylor Encode regions.

column	1	2	3	4	5	6	7
Data source:							
Baylor	■				■	■	■
dbSNP		■		■	■		■
1000 G			■	■		■	■
Number of SNPs:							
chr							
7	779	223	112	87	89	50	141
8	776	209	105	119	85	56	188
9	601	276	101	166	75	33	139
12	472	198	132	195	67	51	171
18	636	227	73	91	95	37	148
2	0	128	168	213	0	0	0
5	409	94	146	131	9	80	76
15	27	428	16	62	1	0	0
16	0	104	209	221	0	0	0
21	362	128	172	190	22	35	48
total	4062	2015	1234	1475	443	342	911
column	1	2	3	4	5	6	7

Column key:	total
1 = Baylor only	4062
2 = dbSNP only	2015
3 = 1000 Genomes only	1234
4 = both dbSNP and 1000 G	1475
5 = both Baylor and dbSNP	443
6 = both Baylor and 1000 G	342
7 = all three sources	911

Fraction of novel SNPs:

Table 7 shows the fraction of submitted SNPs which are novel, versus those already reported in dbSNP build 129. For 1000 Genomes project SNPs as a whole, this fraction is 44.2% novel, 55.8% already in dbSNP. For the 146,000 project SNPs submitted to Illumina, these fractions are 46.1% novel, 53.9% already in dbSNP. To my surprise, the same fractions hold very closely in all five categories of SNPs selected for genotyping, except for the 954 isolated GWAS signals and the Baylor Encode regions. Every GWAS signal must have an rs number from dbSNP in order to appear in the GWAS table. For the ten Baylor Encode regions, the set of SNPs identified both by Baylor and the 1000 Genomes project are only 27% novel; the subset identified only by the 1000 Genomes project is back to 46% novel; and the subset identified only by Baylor – using their much larger set of individuals – shows 90% novel SNPs.

The column in Table 7 labeled 'percent dbSNP only' shows the amount by which each category would have increased if we had submitted all loci from dbSNP in that category, in addition to those found by the 1000 Genomes project. Thus, dbSNP appears to contain 1.86 times as many loci annotated as "missense", "nonsense", "splice-3" or "splice-5" as there are functional SNPs identified in the 1000 Genomes project to date. However, it also fails to annotate 27% of the 1000 Genomes project SNPs which Ensembl does report as functional. The functional annotations from these two sources are less concordant than one would like.

Table 7: Percentage of 1000 Genomes SNPs not found in dbSNP build 129.

category	number submitted	percent of total	percent not in dbSNP	percent already in dbSNP	percent dbSNP only	std. error (percent)	description of category
Comparison of novelty rates across selection categories within the Illumina genotyping panel							
1	55499	36.0	46.3	53.7	132.2	0.2	non-synonymous or splice site SNPs
2	18474	12.0	48.2	51.8	30.0	0.4	four 1 Mb contiguous regions
3	62477	40.6	45.4	54.6	41.3	0.2	100 variable size GWAS regions
4a	954	0.6	0.0	100.0	2.2	–	remaining GWAS isolated peaks
4b	5986	3.9	46.4	53.6	–	0.7	random selection from 1000 Genomes
5	10482	6.8	66.6	33.4	23.8	0.5	Baylor Encode regions, all sources (denominator 8467)
			46.1	53.9		0.1	avg. for all 1000 G calls submitted
Breakdown by source of SNP calls, within the ten Baylor Encode regions							
	2709	1.8	45.6	54.4	–	1.0	1000 Genomes NOT Baylor calls
	1253	0.8	27.3	72.7	–	1.3	1000 Genomes AND Baylor calls
	4505	2.9	90.2	9.8	–	0.5	Baylor calls NOT 1000 Genomes
	8467	5.5	66.6	33.4	–	0.5	union of both sources
	2015	1.3	–	–	23.8	0.5	dbSNP only (submitted)

References:

A large number of web resources have been used in this process, and it is worth recording them here.

Three sets of SNP calls submitted to dbSNP from the 1000 Genomes project:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release ...
/2008_12/CEU.trio.dec.with.x.with.rs.calls.gz
/2008_12/YRI.child.dec.intersect.calls.gz
/2008_12/SOLiD_ReadMe_SNPs_081211.txt # YRI data description
/2008_12/081209.CEU-README.doc # CEU data preliminary description
/2009_02/Pilot1/CEU.snp.gz
/2009_02/Pilot1/JPTCHB.snp.gz
/2009_02/Pilot1/YRI.snp.gz
/2009_02/Pilot1/README_SRP000031_2009_02.txt # low coverage data description

Additional YRI SNP calls from Sanger and AB SOLiD from the following files, after excluding those in the December YRI submission:

ftp://ftp.sanger.ac.uk/pub/1000genomes/F3C-Trio-YRI/DAUGHTER-trio.flt.gz
http://download.solidsoftwaretools/com/misc/misc/na19240.pvalue.out.gz

Description and region boundaries for the Baylor HapMap 3 regions:

http://www.sanger.ac.uk/humgen/hapmap3
http://www.hgsc.bcm.tmc.edu/projects/human

Baylor HapMap 3 SNP calls (this reference from Fu Li Yu at Baylor):
Baylor's sequence data seems not to be available from their ftp site.

ftp://ftp.hgsc.bcm.tmc.edu/pub/data/HapMap3-ENCODE ...
/ENCODE3/ENCODE3v2/bcm-Oct-submission-10202008.txt # Dec 05 21:04 2008
/ENCODE3/ENCODE3v2/Oct_dataRelease_3.doc # Dec 05 21:04 2008

NHGRI GWAS table, last accessed on Wednesday, March 4, 2009, at which time the last line before the table proper stated: "As of 03/03/09, this table includes 273 publications and 1213 SNPs."

http://www.genome.gov/26525384

dbSNP records and chromosome locations from the UCSC Table Browser:

http://genome.ucsc.edu/cgi-bin/hgTables?db=hg18 ...
assembly : Mar.2006
group : Variation and Repeats
track : SNPs (129)
table : snp129 [18 columns x 15,625,346 rows]

List of Ensembl exons:

http://www.ensembl.org/biomart/martview/ ...
Dataset : Homo sapiens genes (NCBI36)
Attributes : Structures / GENE and EXON (everything selected)