Development of a Genotyping Microarray for Studying the Role of Gene-Environment Interactions in Risk for Lung Cancer

Don A. Baldwin,[1,2] Christopher P. Sarnowski,[3] Sabrina A. Reddy,[3] Ian A. Blair,[2,4,5] Margie Clapper,[6] Philip Lazarus,[7] Mingyao Li,[8] Joshua E. Muscat,[9] Trevor M. Penning,[2,4] Anil Vachani,[2,10] and Alexander S. Whitehead[2,4]

[1]Pathonomics LLC, Philadelphia, Pennsylvania 19104, USA; [2]Center of Excellence in Environmental Toxicology, [3]Penn Molecular Profiling Facility, Departments of [4]Pharmacology and [10]Medicine, and Centers for [5]Cancer Pharmacology and [8]Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; [6]Cancer Prevention and Control Program, Fox Chase Cancer Center, Philadelphia, Pennsylvania 19111, USA; [7]Department of Pharmaceutical Sciences, Washington State University, Spokane, Washington 99210, USA; and [9]Department of Public Health Sciences, Pennsylvania State University, Hershey, Pennsylvania 17033, USA

A microarray (LungCaGxE), based on Illumina BeadChip technology, was developed for high-resolution genotyping of genes that are candidates for involvement in environmentally driven aspects of lung cancer oncogenesis and/or tumor growth. The iterative array design process illustrates techniques for managing large panels of candidate genes and optimizing marker selection, aided by a new bioinformatics pipeline component, Tagger Batch Assistant. The LungCaGxE platform targets 298 genes and the proximal genetic regions in which they are located, using ~13,000 DNA single nucleotide polymorphisms (SNPs), which include haplotype linkage markers with a minimum allele frequency of 1% and additional specifically targeted SNPs, for which published reports have indicated functional consequences or associations with lung cancer or other smoking-related diseases. The overall assay conversion rate was 98.9%; 99.0% of markers with a minimum Illumina design score of 0.6 successfully generated allele calls using genomic DNA from a study population of 1873 lung-cancer patients and controls.

Key Words: genetic association, environmental exposures, Tagger Batch Assistant, LungCaGxE

INTRODUCTION

Lung cancer is the leading cause of cancer death for men and women in the United States. The American Cancer Society estimates that in 2013, there will be 228,190 new cases (118,080 in men; 110,110 in women) and 159,480 deaths.[1] Many patients present with disease that is too advanced to treat successfully with surgery and the current portfolio of drugs. Identification of those at highest risk of disease would facilitate earlier diagnosis and therapeutic intervention, with consequent reduced mortality and longer survival time. Risk identification techniques would also support preventative screening and targeted interventions, such as smoking-cessation programs leading to reduced incidence. Given the huge number of new lung cancer cases that occur each year, the impact of such interventions would be significant even if applicable only to an etiologically distinct subset of all cases.

As the majority (up to 90%) of lung cancers occurs in smokers, but only a minority (~10%) of smokers get the disease,[2] it is likely that significant gene/phenotype/environment interactions exist.[3] Although tobacco smoke is the main etiologic agent,[4] the long latency between exposure and disease, the multistep nature of neoplastic transformation,[5] and the low, 10-year lung-cancer risk of elderly, life-long heavy smokers (15%)[6] suggest that factors other than tobacco-associated carcinogens modify risk. These likely include environmental variables,[7] functional genetic polymorphisms,[8,9] and differential expression of genes that interact with such variables.[10]

Strategies to identify associations between genetic variants and diseases, such as lung cancer, include genotyping sequence polymorphisms that are distributed throughout the genome or that occur in specifically targeted genes of interest. Compared with genome-wide approaches, genotyping a focused set of single nucleotide polymorphisms (SNPs) for high-resolution haplotype mapping boosts

Address correspondence to: Alexander S. Whitehead, Perelman School of Medicine, University of Pennsylvania, Room 1311, BRB II/III, 421 Curie Blvd., Philadelphia, PA 19104-6160, USA (Phone: 215-898-2332; E-mail: aswhiteh@exchange.upenn.edu).

doi: 10.7171/jbt.13-2404-004

analysis power for identifying single gene and gene family effects with statistical significance. Targeted, redundant genotyping of candidate genes further enables the analysis of additional variables, such as environmental factors, without a requirement to sample extremely large populations. However, designing a genotyping assay that adequately covers each candidate gene with a sufficiently large number of markers poses a challenge for this approach, especially when interrogating many genes in parallel. Standard genome-wide platforms, such as Affymetrix (Affymetrix, Santa Clara, CA, USA) or Illumina microarrays (Illumina, San Diego, CA, USA), provide predesigned collections of genotyping assays but rarely include enough markers to approach saturation of any given target gene. Microarray vendors therefore offer custom manufacturing options to allow researchers to create comprehensive panels of assays that satisfy the requirements of high-resolution genotyping. We describe a process that connects publicly available SNP catalogs with commercial assay design interfaces, using a new bioinformatics tool that assists with the management of large collections of genes and their haplotype-tagging (HapTag) SNPs. This process was used to demonstrate the rapid and iterative design of a custom-genotyping microarray for studying lung cancer.

MATERIALS AND METHODS

Target Selection

Investigators in our consortium contributed prioritized lists of genes potentially relevant to environmentally mediated biological processes leading to lung cancer. Candidate genes included modulators of and checkpoints within pathways hypothesized to respond to tobacco toxins and environmental factors that may promote oncogenesis, as well as those that may act in concert with environmental factors to support tumor survival, progression, and growth. These genes fell into broad categories, including tobacco-specific nitrosamine [particularly nitrosaminoketone (NNK)] activation and detoxification, polycyclic aromatic hydrocarbon (PAH) activation and detoxification, repair of NNK- and PAH-attributable DNA damage, oxidative stress, inflammatory signaling and processes of immune regulation, steroid hormone metabolism and signaling, nicotine addiction and smoking behavior, and folate transport and metabolism. For each individual gene, HapTag SNPs and genetic polymorphisms known to affect function or shown previously to be associated with risk for lung cancer were sought and if found, incorporated into the final microarray design.

Target sources included extensive literature searches, Ingenuity Pathway Analysis (http://www.ingenuity.com), Database for Annotation, Visualization, and Integrated Discovery (DAVID) Bioinformatics Resources,[11,12] and ongoing research in investigators' laboratories.

SNP Selection

All targeted genes/chromosomal regions were uploaded to the Assay Design Tool (http://support.illumina.com/tools.ilmn; Illumina) for retrieval of all iSelect Infinium database SNPs within each targeted region, as well as from 15 kb sequences flanking the gene-boundary coordinates. Known polymorphisms from the target-selection phase were also queried by reference SNP (rs) number from database of SNPs (dbSNP; http://www.ncbi.nlm.nih.gov/snp/), or uploaded as custom sequences if polymorphisms were unrecognized by iSelect or not annotated in dbSNP. Independently, the targeted genes and regions were analyzed using Tagger (http://www.broadinstitute.org/mpg/tagger/server.html, and International HapMap Project haplotype mapping databases therein)[13] with the following parameters in all combinations: HapMap panels of Utah (U.S.A.) residents of northern and western European ancestry (CEU) and residents of Ibadan, Nigeria of Yoruban ancestry (YRI); SNP minimum allele frequencies 5% and 1%; Tagger mode pairwise and aggressive; SNP *r*2 threshold 0.8; and default settings for all other parameters. The Tagger online interface does not support batch queries using gene symbols, so we created the Tagger Batch Assistant (http://www.bioinformatics.upenn.edu/tagtool/batch.html) as a tool for automated processing of large query lists and management and formatting of the output data.

The retrieved iSelect SNPs were filtered to retain markers with an Infinium design score ≥0.6 (a 60% probability of conversion, i.e., successful genotyping assays for that SNP), and the subset corresponding to selected HapTag SNPs from Tagger was identified. No Infinium design score limits were imposed on functional SNPs from the target selection phase. A panel of 357 ancestry informative markers was included (http://support.illumina.com/array/array_kits/dna_test_panel.ilmn, Illumina catalog GT-17-222).

Genotyping

DNA was extracted from whole-blood samples or buffy-coat fractions using Chemagic DNA purification kits and a Chemagen Magnetic Separation Module I robot (Chemagen/PerkinElmer, Baesweiler, Germany). DNA quality-control checks included A260/280 and E-Gel electrophoresis (Invitrogen, Life Technologies, Grand Island, NY, USA), and DNA samples (n=1873) were normalized to 50 ng/ul and used for genotyping assays. Genotyping was conducted using the iScan system (Illumina), according to the manufacturer's protocols.[14] The Infinium assay amplified and fragmented 200 ng genomic DNA, which was then hybridized to our LungCaGxE iSelect HD Custom

BeadChips containing 24 arrays/BeadChip and 13,308 assayed SNPs/array. Four negative control (no DNA) arrays were processed, and 43 samples were processed twice to check assay consistency. Data from scanned BeadChips were processed in Illumina GenomeStudio for signal quantitation, quality control, and genotype assignments.

The research described does not involve animals. Blood samples from human subjects were collected with their informed consent for research use, including genetic analyses. This study was approved by Institutional Review Boards at the University of Pennsylvania, Pennsylvania State University, Temple University, and Fox Chase Cancer Center.

RESULTS

Tagger Batch Assistant

The online Tagger Batch Assistant tool was designed with two components: one for rapid retrieval of genomic coordinates for large lists of genes and another for managing Tagger output files that result from a batch query using genomic coordinates. Starting with a list of official National Center for Biotechnology Information gene symbols, the tool supports queries of several human genome-build versions, concatenation or separation of overlapping genes, and rules for flanking regions that allow the addition of sequences adjacent to gene coordinates. Multiple choices are available for the amount of flanking sequences added, and rules can be stacked to vary the flanking regions by gene length. The output file can be reviewed in text or spreadsheet formats and is configured for uploading to the Tagger query interface. After receiving compressed Tagger results files, the tool supports automated merging of the user's annotated gene query lists with the corresponding Tagger results.

Assembly of Target Gene Panel

Project investigators identified 298 genes in pathways for which genetically mandated differential interactions with environmental factors leading to lung cancer were deemed to be biologically plausible. These pathways included those supporting or mediating carcinogen effects (i.e., nitrosamine and PAH activation and detoxification), oxidative stress, DNA damage repair, inflammation or immune-system monitoring, estrogen, and other steroid hormone processes, nicotine addiction/smoking behavior, and folate metabolism. Target genes were chosen by examining previous literature, established molecular pathways, and gene interactions and sequence polymorphisms known to affect the functions of genes involved in lung tumor oncogenesis or responses to environmental factors that may impact lung cancer (Table 1). Confirmatory DAVID annotation analyses were performed on the final gene list to summarize the categories represented from Online Mendelian Inheritance in Man (OMIM) Disease, Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway, Gene Ontology (GO) Molecular Function, and GO Biological Process databases (Supplemental Table 1). As expected, the final target panel was confirmed as being enriched for genes associated with risk for lung cancer, folate-sensitive phenotypes, hormone synthesis and signaling, oxidative stress responses, DNA repair, detoxification and metabolism of complex molecules, and apoptosis. Cross-category annotation indicates that the panel is coincidentally enriched for genes involved in schizophrenia, trichothiodystrophy, myocardial infarction, reproductive development, and various neurological processes.

Comparison of Pairwise and Multimarker Tagger Analyses

With the use of dbSNPs for the CEU and YRI populations, Tagger analysis was performed initially to predict marker HapTag SNPs that cover polymorphisms with minimum minor allele frequency (MAF) of 5% and then repeated for MAF >1%. Two Tagger algorithms were compared: pairwise modeling, in which a HapTag marker reports its own genotype and predicts the genotype of one linked SNP, and "aggressive" multimarker modeling, in which the combined genotypes of one to three HapTags report the local haplotype and predict the genotype(s) of one or more linked SNPs.[13,15,16] The resulting number of HapTags calculated for each gene is shown in Table 1. At MAF >1%, pairwise modeling produced a g/h ratio of 1.92 (g=measured+predicted genotypes; h=HapTag markers), and multimarker modeling resulted in 2.38 g/h for the same number of genotypes.

Genotyping Array Design and Assay Performance

Tagger multimarker-predicted HapTags with MAF >1% were filtered for iSelect Infinium design scores ≥0.6. *TLR5* had no multimarker HapTags, so pairwise HapTags were selected; *CCR2*, *UGT2B15*, and *GSTT1* had no HapTags, so marker SNPs were manually identified. To avoid exceeding the marker capacity set by our microarray manufacturing budget, the low-priority genes, *ALPL*, *TNS1*, *GAB1*, *HHIP*, *DBH*, and *PTGIS*, were dropped, and HapTag coverage of *GPR126* was reduced to 85%. With the addition of specifically targeted functional SNPs and published marker SNPs, 12,890 genomic SNPs were compiled for the final design of the LungCaGxE array with average and median intermarker distances of 5958 bp and 1093 bp, respectively. Sixty-one mitochondrial DNA SNPs were included to target *MT-COI*, as well as 357 ancestry informative markers for a total of 13,308 genotyping markers on

TABLE 1

Targeted Genes, Annotations, and Number of HapTag SNPs Identified by Pairwise or Multimarker (multi) Algorithms at the Indicated Minor Allele Frequencies (MAFs) and Infinium (Inf) Design Scores ≥0.6

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
A2BP1		1099	1046	914	710	599	full	ataxin 2 binding protein 1	tum
ABCB1		91	81	62	62	44	full	ATP binding cassette, subfamily B [multidrug resistance (MDR)/transporter associated with antigen processing (TAP)], member 1	tum
ABCC1		199	180	162	142	123	full	ATP binding cassette, subfamily C [cystic fibrosis transmembrane conductance regulator (CFTR)/multidrug resistance-associated protein (MRP)], member 1	mut/PAH
ABCC2		51	44	33	40	29	full	ATP binding cassette, subfamily C (CFTR/MRP), member 2	fol
ABCC4		355	324	274	244	199	full	ATP binding cassette, subfamily C (CFTR/MRP), member 4	mut/PAH
ACHE		12	12	10	12	10	full	ACETYLCHOLINESTERASE (YT BLOOD GROUP)	nic
ADAM19		110	101	78	81	61	full	a disintegrin and metalloprotease domain (ADAM) metallopeptidase domain 19 (meltrin β)	adh
ADH1B		28	26	23	23	20	full	alcohol dehydrogenase 1B (class I), β polypeptide	tox
ADH7		49	47	40	40	33	full	alcohol dehydrogenase 7 (class IV), μ or σ polypeptide	tox
ADK		120	98	80	74	55	full	adenosine kinase	inf
AGER	PPT2	23	22	20	22	17	full	advanced glycosylation end product-specific receptor	inf/mut
AHCY		17	12	12	8	8	full	S-ADENOSYLHOMOCYSTEINE HYDROLASE	fol
AHR		27	27	24	25	21	full	ARYL-HYDROCARBON RECEPTOR	PAH
AHRR		87	79	71	61	51	full	ARYL-HYDROCARBON RECEPTOR REPRESSOR	PAH
AKAP9		23	21	17	19	15	full	A kinase (PRKA) anchor protein (yotiao) 9	tum
AKR1A1		17	13	10	12	7	full	ALDO-KETO REDUCTASE FAMILY 1, MEMBER A1 (ALDEHYDE REDUCTASE)	PAH
AKR1B10		30	25	24	22	21	full	ALDO-KETO REDUCTASE FAMILY 1, MEMBER B10 (ALDOSE REDUCTASE-LIKE)	onc

Continued

TABLE 1

(*Continued*)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
AKR1C1	AKR1C2	18	15	15	22	21	full	ALDO-KETO REDUCTASE FAMILY 1, MEMBER C1 [DIHYDRODIOL DEHYDROGENASE 1; 20-α (3-α)-HYDROXYSTEROID DEHYDROGENASE]	nit/PAH/ str
AKR1C2	AKR1C1	46	38	36	24	23	full	ALDO-KETO REDUCTASE FAMILY 1, MEMBER C2 (DIHYDRODIOL DEHYDROGENASE 2; BILE ACID BINDING PROTEIN; 3-α HYDROXYSTEROID DEHYDROGENASE, TYPE III)	nit/PAH/ str
AKR1C3		44	36	35	27	26	full	ALDO-KETO REDUCTASE FAMILY 1, MEMBER C3 (3-α HYDROXYSTEROID DEHYDROGENASE, TYPE II)	PAH/str
AKT1		16	13	12	13	12	full	V-AKT MURINE THYMOMA VIRAL ONCOGENE HOMOLOG 1	onc
AKT2		15	14	11	13	10	full	v-akt murine thymoma viral oncogene homolog 2	tum
AKT3		95	89	70	69	54	full	v-akt murine thymoma viral oncogene homolog 3 (PKB, γ)	tum
ALDH1L1		115	98	89	67	58	full	aldehyde dehydrogenase 1 family, member L1	fo1
ALOX5		65	57	48	46	36	full	ARACHIDONATE 5-LIPOXYGENASE	inf/oxs
ALPL		100	96	88	79	70	dropped for capacity	alkaline phosphatase, liver/bone/kidney	
ANKK1	DRD2	44	41	33	26	19	full	ANKYRIN REPEAT AND KINASE DOMAIN CONTAINING 1	nic
APEX1		30	29	26	26	22	full	APEX NUCLEASE (MULTIFUNCTIONAL DNA REPAIR ENZYME) 1	DNA
AR		15	5	5	12	12	full	ANDROGEN RECEPTOR (DIHYDROTESTOSTERONE RECEPTOR; TESTICULAR FEMINIZATION; SPINAL AND BULBAR MUSCULAR ATROPHY; KENNEDY DISEASE)	str
AREG		31	14	12	10	7	full	AMPHIREGULIN (SCHWANNOMA-DERIVED GROWTH FACTOR)	onc
ARID1A		10	10	7	10	7	full	AT RICH-INTERACTIVE DOMAIN 1A (SWI-LIKE)	onc
ARNT		30	29	17	27	15	full	ARYL-HYDROCARBON RECEPTOR NUCLEAR TRANSLOCATOR	PAH

Continued

TABLE 1

(*Continued*)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
ARNTL		103	98	88	84	75	full	ARYL-HYDROCARBON RECEPTOR NUCLEAR TRANSLOCATOR-LIKE	PAH
ATIC		38	34	31	28	25	full	5-AMINOIMIDAZOLE-4-CARBOXAMIDE RIBONUCLEOTIDE FORMYLTRANSFERASE/ IMP CYCLOHYDROLASE	fol
BCL2		229	224	178	190	144	full	B CELL chronic lymphocytic leukemia (CLL)/ LYMPHOMA 2	onc
BDNF		33	33	26	30	23	full	BRAIN-DERIVED NEUROTROPHIC FACTOR	nic
BHMT		28	28	25	26	22	full	BETAINE-HOMOCYSTEINE METHYLTRANSFERASE	fol
BIRC5		21	19	18	17	16	full	BACULOVIRAL inhibitor of apoptosis (IAP) REPEAT-CONTAINING 5 (SURVIVIN)	onc
BMPR1B		210	199	180	143	125	full	BONE MORPHOGENETIC PROTEIN RECEPTOR, TYPE IB	onc
BRCA2		88	82	59	68	47	full	breast cancer 2, early onset	tum
C3		63	59	49	55	45	full	COMPLEMENT COMPONENT 3	inf
CAMKK1		41	36	33	35	32	full	calcium/calmodulin-dependent protein kinase kinase 1, α	tum
CBR1		19	17	15	13	11	full	CARBONYL REDUCTASE 1	nit/PAH
CBR3		16	13	11	11	9	full	CARBONYL REDUCTASE 3	nit/PAH
CBS		60	57	53	47	43	full	CYSTATHIONINE-β-SYNTHASE	fol
CCL2		23	20	17	16	13	full	chemokine (C–C motif) ligand 2	inf
CCL21		20	18	16	15	14	full	chemokine (C–C motif) ligand 21	inf
CCL5		10	9	6	8	5	full	chemokine (C–C motif) ligand	inf
CCNA2		22	17	12	13	9	full	cyclin A2	onc
CCND1		25	25	20	23	18	full	CYCLIN D1	onc
CCND3		83	21	17	18	14	full	cyclin D3	onc
CCR2		6	0	0	0	0	nine non-HapTag SNPs	chemokine (C–C motif) receptor 2	inf
CD47		46	43	35	36	28	full	CD47 ANTIGEN (RH-RELATED ANTIGEN, INTEGRIN-ASSOCIATED SIGNAL TRANSDUCER)	adh
CDH1		66	56	53	50	47	full	CADHERIN 1, TYPE 1, E-CADHERIN (EPITHELIAL)	adh

Continued

TABLE 1

(*Continued*)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
CDKN2A		33	31	27	27	24	full	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	onc
CES3		10	8	5	7	4	full	CARBOXYLESTERASE 3	tox
CHAT	SLC18A3	85	98	81	112	93	full	CHOLINE ACETYLTRANSFERASE	nic
CHRNA3	CHRNA5	2	2	1	22	18	full	CHOLINERGIC RECEPTOR, NICOTINIC, α 3	nic
CHRNA4		28	25	24	21	20	full	CHOLINERGIC RECEPTOR, NICOTINIC, α 4	nic
CHRNA5	CHRNB4	28	26	21	13	10	full	CHOLINERGIC RECEPTOR, NICOTINIC, α 5	nic
CHRNA7		89	85	71	76	62	full	CHOLINERGIC RECEPTOR, NICOTINIC, α 7	nic
CHRNB2		24	23	22	20	18	full	CHOLINERGIC RECEPTOR, NICOTINIC, β 2 (NEURONAL)	nic
CHRNB3		21	20	17	16	13	full	CHOLINERGIC RECEPTOR, NICOTINIC, β 3	nic
CHRNB4	CHRNA3	27	26	24	15	14	full	CHOLINERGIC RECEPTOR, NICOTINIC, β 4	nic
CHUK		27	26	20	23	17	full	CONSERVED HELIX-LOOP-HELIX UBIQUITOUS KINASE	onc
CLOCK		35	29	27	20	18	full	CLOCK HOMOLOG (MOUSE)	onc
COL3A1		55	54	46	46	37	full	COLLAGEN, TYPE III, α 1 (EHLERS-DANLOS SYNDROME TYPE IV, AUTOSOMAL DOMINANT)	adh
COMT		56	49	43	39	34	full	CATECHOL-O-METHYLTRANSFERASE	PAH/str
CRP		31	31	23	25	18	full	C-REACTIVE PROTEIN, PENTRAXIN-RELATED	inf
CRY1		51	45	38	36	28	full	CRYPTOCHROME 1 (PHOTOLYASE-LIKE)	onc
CSNK1D		19	15	11	15	11	full	CASEIN KINASE 1, δ	onc
CTH		38	34	31	30	27	full	CYSTATHIONASE (CYSTATHIONINE γ-LYASE)	fol
CTLA4		27	27	24	18	15	full	CYTOTOXIC T-LYMPHOCYTE-ASSOCIATED PROTEIN 4	inf
CTSD		9	8	7	8	7	full	CATHEPSIN D (LYSOSOMAL ASPARTYL PEPTIDASE)	tum
CYP17A1		24	19	15	16	12	full	CYTOCHROME P450, FAMILY 17, SUBFAMILY A, POLYPEPTIDE 1	str
CYP19A1		112	104	94	75	65	full	CYTOCHROME P450, FAMILY 19, SUBFAMILY A, POLYPEPTIDE 1	str
CYP1A1	CYP1A2	13	11	11	11	10	full	CYTOCHROME P450, FAMILY 1, SUBFAMILY A, POLYPEPTIDE 1	PAH

Continued

TABLE 1

(*Continued*)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
CYP1A2	CYP1A1	13	13	10	11	9	full	CYTOCHROME P450, FAMILY 1, SUBFAMILY A, POLYPEPTIDE 2	PAH
CYP1B1		41	38	34	33	28	full	CYTOCHROME P450, FAMILY 1, SUBFAMILY B, POLYPEPTIDE 1	PAH
CYP21A2		15	7	5	7	5	full	CYTOCHROME P450, FAMILY 21, SUBFAMILY A, POLYPEPTIDE 2	str
CYP2A13		11	7	6	7	6	full	cytochrome P450, family 2, subfamily A, polypeptide 13	nit
CYP2A6		18	13	11	13	11	full	CYTOCHROME P450, FAMILY 2, SUBFAMILY A, POLYPEPTIDE 6	nit
CYP2B6		43	34	31	27	23	full	cytochrome P450, family 2, subfamily B, polypeptide 6	nit/tum
CYP2C9		32	28	20	26	18	full	CYTOCHROME P450, FAMILY 2, SUBFAMILY C, POLYPEPTIDE 9	PAH
CYP2D6		11	5	5	5	5	full	CYTOCHROME P450, FAMILY 2, SUBFAMILY D, POLYPEPTIDE 6	nit
CYP2E1		39	36	35	30	29	full	cytochrome P450, family 2, subfamily E, polypeptide 1	nit
CYP3A4		30	24	18	23	17	full	CYTOCHROME P450, SUBFAMILY IIIA (NIPHEDIPINE OXIDASE), POLYPEPTIDE 3	str
DBH		88	85	75	73	63	dropped for capacity	DOPAMINE β-HYDROXYLASE (DOPAMINE β-MONOOXYGENASE)	
DDX54		10	10	9	9	8	full	DEAD (ASP-GLU-ALA-ASP) BOX POLYPEPTIDE 54	onc
DHFR		24	19	16	15	12	full	DIHYDROFOLATE REDUCTASE	fol
DMGDH		75	69	64	58	50	full	dimethylglycine dehydrogenase	fol
DNMT1		18	16	12	14	10	full	DNA (cytosine-5-)-methyltransferase 1	fol
DNMT3A		56	54	44	45	35	full	DNA (cytosine-5-)-methyltransferase 3 α	fol
DNMT3B		58	56	46	40	30	full	DNA (cytosine-5-)-methyltransferase 3 β	fol
DRD2	ANKK1	53	53	45	44	36	full	DOPAMINE RECEPTOR D2	nic
DRD4		11	11	10	10	9	full	DOPAMINE RECEPTOR D4	nic
EGF		129	115	79	95	61	full	epidermal growth factor	tum

Continued

TABLE 1

(Continued)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
EGFR		212	196	167	162	135	full	EPIDERMAL GROWTH FACTOR RECEPTOR (ERYTHROBLASTIC LEUKEMIA VIRAL (V-ERB-B) ONCOGENE HOMOLOG, AVIAN)	onc
EGLN2		22	20	18	17	15	full	egl nine homolog 2	oxs
EPHX1		36	26	23	26	22	full	EPOXIDE HYDROLASE 1, MICROSOMAL (XENOBIOTIC)	PAH
ERCC1		20	18	14	16	12	full	EXCISION REPAIR CROSS-COMPLEMENTING RODENT REPAIR DEFICIENCY, COMPLEMENTATION GROUP 1 (INCLUDES OVERLAPPING ANTISENSE SEQUENCE)	DNA
ERCC2		34	33	22	30	20	full	EXCISION REPAIR CROSS-COMPLEMENTING RODENT REPAIR DEFICIENCY, COMPLEMENTATION GROUP 2 (XERODERMA PIGMENTOSUM D)	DNA
ERCC3		39	36	25	32	20	full	excision repair cross-complementing rodent repair deficiency, complementation group 3 (xeroderma pigmentosum group B complementing)	DNA
ERCC4		61	55	39	41	26	full	EXCISION REPAIR CROSS-COMPLEMENTING RODENT REPAIR DEFICIENCY, COMPLEMENTATION GROUP 4	DNA
ERCC5		67	58	42	46	30	full	EXCISION REPAIR CROSS-COMPLEMENTING RODENT REPAIR DEFICIENCY, COMPLEMENTATION GROUP 5 [XERODERMA PIGMENTOSUM, COMPLEMENTATION GROUP G (COCKAYNE SYNDROME)]	DNA
ERCC6		64	58	37	46	25	full	excision repair cross-complementing rodent repair deficiency, complementation group 6	DNA
ERCC8		38	33	21	28	17	full	EXCISION REPAIR CROSS-COMPLEMENTING RODENT REPAIR DEFICIENCY, COMPLEMENTATION GROUP 8	DNA
ESR1		341	237	173	182	126	full	ESTROGEN RECEPTOR 1	str
ESR2		68	61	52	43	34	full	ESTROGEN RECEPTOR 2 (ER β)	str
EYA2		342	325	293	247	217	full	eyes absent homolog 2 (Drosophila)	DNA

Continued

TABLE 1

(*Continued*)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
FAM13A		165	156	138	113	93	full	family with sequence similarity 13, member A	mut
FCGR1A		8	1	1	1	1	full	Fc fragment of IgG, high-affinity Ia, receptor (CD64)	inf
FKBP5		53	30	22	24	18	full	FK506 BINDING PROTEIN 5	inf/str
FMO3		46	42	34	34	26	full	Flavin containing monooxygenase 3	tox
FOLH1		22	14	10	13	9	full	FOLATE HYDROLASE (PROSTATE-SPECIFIC MEMBRANE ANTIGEN) 1	fol
FOLR1	FOLR2	5	5	4	8	6	full	FOLATE RECEPTOR 1 (ADULT)	fol
FOLR2	FOLR1	9	7	6	3	3	full	FOLATE RECEPTOR 2 (FETAL)	fol
FOLR3		23	13	12	9	9	full	FOLATE RECEPTOR 3 (γ)	fol
FPGS		23	20	17	19	16	full	FOLYLPOLYGLUTAMATE SYNTHASE	fol
FTCD		51	50	46	43	39	full	formiminotransferase cyclodeaminase	fol
GAB1		72	68	58	56	46	dropped for capacity	growth factor receptor-bound protein 2-associated binding protein 1	
GART		22	19	14	19	14	full	PHOSPHORIBOSYLGLYCINAMIDE FORMYLTRANSFERASE, PHOSPHORIBOSYLGLYCINAMIDE SYNTHETASE, PHOSPHORIBOSYLAMINOIMIDAZOLE SYNTHETASE	fol
GATA3		63	62	50	58	45	full	GATA BINDING PROTEIN 3	tum
GCLC		85	76	69	66	59	full	GLUTAMATE-CYSTEINE LIGASE, CATALYTIC SUBUNIT	oxs
GCLM		17	16	15	15	12	full	GLUTAMATE-CYSTEINE LIGASE, MODIFIER SUBUNIT	oxs
GDF15		18	16	15	15	14	full	growth differentiation factor 15	onc
GGH		27	21	13	16	9	full	γ-glutamyl hydrolase (conjugase, folylpolygammaglutamyl hydrolase)	fol
GHR		93	84	72	66	55	full	growth hormone receptor	tum
GNMT		17	16	13	15	12	full	glycine N-methyltransferase	fol
GPC5		969	889	739	666	528	full	glypican 5	mut
GPER		15	12	11	12	11	full	GPCR 30	str
GPR126		70	63	52	54	44	85%	GPCR 126	adh

Continued

TABLE 1

(*Continued*)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
GPX1		14	10	8	10	8	full	GLUTATHIONE PEROXIDASE 1	oxs
GPX3		53	51	47	43	39	full	GLUTATHIONE PEROXIDASE 3 (PLASMA)	oxs
GRPR		13	13	13	22	22	full	GASTRIN-RELEASING PEPTIDE RECEPTOR	onc
GSK3B		71	58	44	48	35	full	glycogen synthase kinase 3 β	tum
GSR		49	42	32	37	27	full	GLUTATHIONE REDUCTASE	oxs
GSS		23	20	19	18	17	full	GLUTATHIONE SYNTHETASE	fol
GSTA1		16	11	9	10	8	full	GLUTATHIONE S-TRANSFERASE A1	oxs/PAH
GSTA4		37	34	24	27	17	full	GLUTATHIONE S-TRANSFERASE A4	oxs
GSTCD		39	37	27	29	18	full	glutathione S-transferase, C-terminal domain containing	fol
GSTM1	GSTM2	7	4	6	7	7	full	GLUTATHIONE S-TRANSFERASE M1	oxs/PAH
GSTM2	GSTM1	15	12	10	10	8	full	GLUTATHIONE S-TRANSFERASE M2 (MUSCLE)	oxs/PAH
GSTM5	GSTM1	25	14	12	11	10	full	GLUTATHIONE S-TRANSFERASE M5	oxs
GSTO1		33	30	24	28	23	full	GLUTATHIONE S-TRANSFERASE ω 1	oxs
GSTP1		21	19	17	16	14	full	GLUTATHIONE S-TRANSFERASE π 1	oxs/PAH
GSTT1		1	0	0	0	0	four non-HapTag SNPs	GLUTATHIONE S-TRANSFERASE θ 1	oxs/PAH
HDC		38	36	33	32	29	full	HISTIDINE DECARBOXYLASE	inf
HELQ		40	37	31	26	20	full	HELQ helicase, POLQ-like	DNA
HFE		24	23	20	19	17	full	HEMOCHROMATOSIS	tox
HGF		83	78	59	64	46	full	HEPATOCYTE GROWTH FACTOR (HEPAPOIETIN A; SCATTER FACTOR)	onc
HHIP		43	43	37	35	28	dropped for capacity	Hedgehog-interacting protein	
hsa-mir21		97					full	HOMO SAPIENS MICRORNA 21	onc
HSD11B1		39	37	35	29	27	full	HYDROXYSTEROID (11-β) DEHYDROGENASE 1	nit/str
HSD17B1		17	14	12	13	11	full	HYDROXYSTEROID (17-β) DEHYDROGENASE 1	str
HSD17B12		84	77	63	64	51	full	HYDROXYSTEROID (17-β) DEHYDROGENASE 12	str
HSD17B3		63	56	46	45	35	full	HYDROXYSTEROID (17-β) DEHYDROGENASE 3	str
HSD17B7		24	20	19	17	16	full	HYDROXYSTEROID (17-β) DEHYDROGENASE 7	str
HSD3B1		20	16	16	11	11	full	HYDROXY-δ-5-STEROID DEHYDROGENASE, 3 β- AND STEROID δ-ISOMERASE 1	str

Continued

TABLE 1

(*Continued*)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
HTR3E		24	22	18	20	16	full	5-hydroxytryptamine (serotonin) receptor 3, family member E	nic
HTR4		126	115	105	89	81	full	5-hydroxytryptamine (serotonin) receptor 4	nic
ICAM1		26	25	21	25	21	full	intercellular adhesion molecule 1	inf
ID2		13	13	11	12	11	full	INHIBITOR OF DNA BINDING 2, DOMINANT NEGATIVE HELIX-LOOP-HELIX PROTEIN	onc
IDH1		35	31	24	27	22	full	ISOCITRATE DEHYDROGENASE 1 (NADP+), SOLUBLE	onc
IER3		21	19	18	16	15	full	immediate early response 3	mut/inf
IFNG		40	38	30	31	23	full	IFN-γ	inf
IGF1		65	57	38	48	30	full	insulin-like growth factor 1 (somatomedin C)	onc
IGF1R		318	306	272	239	206	full	insulin-like growth factor 1 receptor	onc
IGF2	TH	16	14	16	17	16	full	insulin-like growth factor 2 (somatomedin A)	onc
IGF2R		161	148	108	123	86	full	insulin-like growth factor 2 receptor	onc
IGFBP3		24	19	18	19	18	full	INSULIN-LIKE GROWTH FACTOR BINDING PROTEIN 3	onc
IKBKB		25	24	18	20	15	full	inhibitor of κ light polypeptide gene enhancer in B-cells, kinase β	inf
IL10		34	30	24	28	21	full	INTERLEUKIN 10	inf
IL1B		25	24	20	23	19	full	interleukin 1, β	inf
IL1RN		66	66	58	54	46	full	interleukin 1 receptor antagonist	inf
IL4		49	46	40	39	33	full	INTERLEUKIN 4	inf
IL6		44	39	32	34	26	full	INTERLEUKIN 6	inf
IL8		30	28	24	21	17	full	interleukin 8	inf
IRS1		46	43	33	37	28	full	INSULIN RECEPTOR SUBSTRATE 1	onc
JUN		19	18	16	17	15	full	jun oncogene	onc
KEAP1		11	9	9	9	9	full	kelch-like ECH-associated protein 1	oxs
KLRK1		27	21	20	19	18	full	KILLER CELL LECTIN-LIKE RECEPTOR SUBFAMILY C, MEMBER 4	adh
KRT18		15	10	7	10	7	full	KERATIN 18	adh
KRT19		19	19	18	17	16	full	KERATIN 19	adh
LTA	TNF	10	9	10	19	10	full	lymphotoxin α (TNF superfamily, member 1)	inf
LTC4S		6	5	5	5	5	full	LEUKOTRIENE C4 SYNTHASE	inf
MAF		48	47	43	45	41	full	v-maf musculoaponeurotic fibrosarcoma oncogene homolog (avian)	onc

Continued

TABLE 1

(*Continued*)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
MAOA		19	19	19	22	22	full	MONOAMINE OXIDASE A	nic
MDM2		40	31	24	27	20	full	MDM2, TRANSFORMED 3T3 CELL DOUBLE-MINUTE 2, P53 BINDING PROTEIN (MOUSE)	DNA
MGMT		249	238	212	179	152	full	O-6-METHYLGUANINE-DNA METHYLTRANSFERASE	DNA/nit
MGST3		62	56	53	48	45	full	MICROSOMAL GST 3	oxs
MIF		44	42	39	35	31	full	macrophage migration inhibitory factor (glycosylation-inhibiting factor)	inf
MSR1		99	94	81	79	68	full	macrophage scavenger receptor 1	inf
MTAP		64	59	52	44	37	full	methylthioadenosine phosphorylase	fol
MT-COI							61 SNPs	mitochondrially encoded cytochrome c oxidase I	oxs
MTHFD1		36	32	29	31	28	full	METHYLENETETRAHYDROFOLATE DEHYDROGENASE (NADP+ DEPENDENT) 1, METHENYLTETRAHYDROFOLATE CYCLOHYDROLASE, FORMYLTETRAHYDROFOLATE SYNTHETASE	fol
MTHFR		36	32	29	26	22	full	5,10-METHYLENETETRAHYDROFOLATE REDUCTASE (NADPH)	fol
MTHFS		57	53	47	35	29	full	5,10-METHENYLTETRAHYDROFOLATE SYNTHETASE (5-FORMYLTETRAHYDROFOLATE CYCLO-LIGASE)	fol
MTR		79	64	49	42	33	full	5-METHYLTETRAHYDROFOLATE-HOMOCYSTEINE METHYLTRANSFERASE	fol
MTRR		54	49	36	34	22	full	5-METHYLTETRAHYDROFOLATE-HOMOCYSTEINE METHYLTRANSFERASE REDUCTASE	fol
MUTYH		17	17	14	16	13	full	MUTY HOMOLOG (*Escherichia coli*)	DNA/oxs
MYBL2		34	32	26	24	18	full	v-myb myeloblastosis viral oncogene homolog (avian)-like 2	tum
MYC		40	38	27	36	24	full	V-MYC MYELOCYTOMATOSIS VIRAL ONCOGENE HOMOLOG (AVIAN)	onc

Continued

TABLE 1

(*Continued*)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
NAT1		115	99	86	80	68	full	N-ACETYLTRANSFERASE 1 (ARYLAMINE N-ACETYLTRANSFERASE)	fol
NAT2		56	50	43	39	32	full	N-ACETYLTRANSFERASE 2 (ARYLAMINE N-ACETYLTRANSFERASE)	PAH
NCOA6		19	17	16	14	13	full	NUCLEAR RECEPTOR COACTIVATOR 6	onc
NFE2L2		30	30	24	28	21	full	NUCLEAR FACTOR (ERYTHROID-DERIVED 2)-LIKE 2	oxs
NFKB1		74	72	54	56	39	full	NUCLEAR FACTOR OF κ LIGHT POLYPEPTIDE GENE ENHANCER IN B CELLS 1 (P105)	inf/onc
NFKBIA		27	26	25	23	21	full	NUCLEAR FACTOR OF κ LIGHT POLYPEPTIDE GENE ENHANCER IN B CELLS INHIBITOR, α	inf/onc
NOS1		191	172	141	137	108	full	NITRIC OXIDE SYNTHASE 1 (NEURONAL)	inf
NOS2		56	53	51	47	44	full	NITRIC OXIDE SYNTHASE 2A (INDUCIBLE, HEPATOCYTES)	inf
NOS3		30	29	23	26	20	full	NITRIC OXIDE SYNTHASE 3 (ENDOTHELIAL CELL)	inf
NQO1		20	19	19	15	15	full	NAD(P)H DEHYDROGENASE, QUINONE 1	oxs/PAH
NR1D2		27	22	18	21	17	full	NUCLEAR RECEPTOR SUBFAMILY 1, GROUP D, MEMBER 2	onc
NR3C1		58	56	43	48	36	full	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)	onc
NRIP1		85	33	30	30	26	full	NUCLEAR RECEPTOR-INTERACTING PROTEIN 1	onc
NSD1		46	33	26	27	20	full	NUCLEAR RECEPTOR BINDING SET DOMAIN PROTEIN 1	onc
OAS1		34	30	26	26	22	full	2′,5′-oligoadenylate synthetase 1, 40/46 kDa	inf
OAS2		62	51	46	38	32	full	2′,5′-oligoadenylate synthetase 2, 69/71 kDa	inf
OGG1		19	18	14	17	12	full	8-OXOGUANINE DNA GLYCOSYLASE	DNA/oxs
OPRM1		209	174	149	139	115	full	OPIOID RECEPTOR, μ 1	nic
PCDH7		179	174	143	135	109	full	protocadherin 7	adh
PER1		14	14	11	14	11	full	PERIOD HOMOLOG 1 (DROSOPHILA)	onc
PGR		91	71	56	53	39	full	PROGESTERONE RECEPTOR	str
PHB2		9	9	8	9	8	full	PROHIBITIN 2	adh/str
PID1		234	221	200	173	153	full	Phosphotyrosine-interaction domain containing 1	inf

Continued

TABLE 1

(*Continued*)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
PIK3CG		47	45	36	31	22	full	phosphoinositide-3-kinase, catalytic, γ polypeptide	onc
PLA2G6		57	55	43	42	30	full	phospholipase A2, group VI (cytosolic, calcium-independent)	tum
PLEKHA6		134	126	120	90	85	full	pleckstrin homology domain containing, family A member 6	nic
POLH		17	14	12	14	11	full	POLYMERASE (DNA-DIRECTED), η	DNA
POLI		19	18	10	18	10	full	POLYMERASE (DNA-DIRECTED) ι	DNA
POLK		39	35	26	34	25	full	POLYMERASE (DNA-DIRECTED) κ	DNA
POLL		20	19	15	18	13	full	POLYMERASE (DNA-DIRECTED), λ	DNA
PON1		67	66	59	56	48	full	paraoxonase 1	tox
PPARG		113	107	80	83	60	full	PEROXISOME PROLIFERATIVE ACTIVATED RECEPTOR, γ	onc
PPARGC1B		150	141	132	112	102	full	PEROXISOME PROLIFERATIVE ACTIVATED RECEPTOR, γ, COACTIVATOR 1, β	onc
PPT2	AGER	18	18	14	15	13	full	palmitoyl-protein thioesterase 2	tox
PTCH1		20	20	20	18	18	full	patched homolog 1	tum
PTEN		30	28	20	27	18	full	phosphatase and tensin homolog	onc
PTGIS		65	59	47	57	44	dropped for capacity	PG I2 (prostacyclin) synthase	
PTGS1		73	70	54	61	47	full	PG-endoperoxide synthase 1 (PG G/H synthase and cyclooxygenase)	inf
PTGS2		29	24	19	19	14	full	PG-ENDOPEROXIDE SYNTHASE 2 (PG G/H SYNTHASE AND COX)	infl/oxs
RELA		13	13	11	12	10	full	v-rel reticuloendotheliosis viral oncogene homolog A (avian)	onc
RERGL		23	19	15	17	13	full	RAS-like, estrogen-regulated, growth inhibitor (RERG)/RAS-like	str
RNASEL		27	26	23	25	22	full	ribonuclease L (2′,5′-oligoisoadenylate synthetase-dependent)	inf
SELE		49	45	32	38	25	full	selectin E	inf
SERPINA3		40	39	37	37	34	full	SERPIN PEPTIDASE INHIBITOR, CLADE A (α-1 ANTIPROTEINASE, ANTITRYPSIN), MEMBER 3	adh/onc

Continued

TABLE 1

(*Continued*)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
SHMT1		29	23	20	18	14	full	SERINE HYDROXYMETHYLTRANSFERASE 1 (SOLUBLE)	fol
SHMT2		8	6	6	6	6	full	SERINE HYDROXYMETHYLTRANSFERASE 2 (MITOCHONDRIAL)	fol
SLC18A3	CHAT	16	32	36	6	10	full	SOLUTE CARRIER FAMILY 18 (VESICULAR ACETYLCHOLINE), MEMBER 3	nic
SLC19A1		31	29	26	25	22	full	SOLUTE CARRIER FAMILY 19 (FOLATE TRANSPORTER), MEMBER 1	fol
SLC5A7		46	45	36	38	29	full	SOLUTE CARRIER FAMILY 5 (CHOLINE TRANSPORTER), MEMBER 7	nic
SLC6A3		55	47	43	41	37	full	SOLUTE CARRIER FAMILY 6 (NEUROTRANSMITTER TRANSPORTER, DOPAMINE), MEMBER 3	nic
SLC7A5		58	51	46	44	40	full	SOLUTE CARRIER FAMILY 7 (CATIONIC AMINO ACID TRANSPORTER, Y+ SYSTEM), MEMBER 5	onc
SOD1		17	15	14	15	14	full	SUPEROXIDE DISMUTASE 1, SOLUBLE [AMYOTROPHIC LATERAL SCLEROSIS 1 (ADULT)]	inf
SOD2		17	15	13	13	11	full	SUPEROXIDE DISMUTASE 2, MITOCHONDRIAL	oxs
SOD3		25	21	18	20	16	full	SUPEROXIDE DISMUTASE 3, EXTRACELLULAR	inf
STC2		30	27	26	24	23	full	STANNIOCALCIN 2	onc
SULT1A1		14	10	7	10	7	full	SULFOTRANSFERASE FAMILY, CYTOSOLIC, 1A, PHENOL-PREFERRING, MEMBER 1	PAH
SULT1E1		54	49	28	43	22	full	SULFOTRANSFERASE FAMILY 1E, ESTROGEN-PREFERRING, MEMBER 1	str
SULT2A1		28	26	24	21	18	full	SULFOTRANSFERASE FAMILY, CYTOSOLIC, 2A, DEHYDROEPIANDROSTERONE (DHEA)-PREFERRING, MEMBER 1	str
TCN2		41	40	38	33	29	full	TRANSCOBALAMIN II; MACROCYTIC ANEMIA	fol
TEF		32	21	20	18	17	full	THYROTROPHIC EMBRYONIC FACTOR	onc

Continued

TABLE 1

(*Continued*)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
TFF1		72	64	57	49	43	full	TREFOIL FACTOR 1 (BREAST CANCER, ESTROGEN-INDUCIBLE SEQUENCE EXPRESSED IN)	onc
TFF3		46	44	39	39	35	full	TREFOIL FACTOR 3 (INTESTINAL)	onc
TGFA		136	129	107	109	87	full	TRANSFORMING GROWTH FACTOR, α	onc
TGFB1		19	19	17	18	16	full	TRANSFORMING GROWTH FACTOR, β 1 (CAMURATI-ENGELMANN DISEASE)	onc
TGFBR1		37	34	24	29	19	full	TRANSFORMING GROWTH FACTOR, β RECEPTOR I (ACTIVIN A RECEPTOR TYPE II-LIKE KINASE, 53 KDA)	onc
TH	IGF2	28	26	20	20	18	full	TYROSINE HYDROXYLASE	nic
THSD4		594	558	498	428	364	full	thrombospondin, type I, domain containing 4	adh/inf
TLR1	TLR6	20	20	17	23	19	full	Toll-like receptor 1	inf
TLR10		35	30	24	21	17	full	Toll-like receptor 10	inf
TLR2		23	20	20	19	19	full	Toll-like receptor 2	inf
TLR4		58	56	43	50	38	full	Toll-like receptor 4	inf
TLR5		14	12	12	0	0	13 pairwise HapTags	Toll-like receptor 5	inf
TLR6	TLR1	32	12	9	9	6	full	Toll-like receptor 6	inf
TNF	LTA	20	17	6	4	3	full	tumor necrosis factor	inf
TNS1		154	151	133	135	114	dropped for capacity	tensin 1	
TP53		19	17	17	15	15	full	TUMOR PROTEIN P53 (LI-FRAUMENI SYNDROME)	onc
TP53BP1		25	23	18	20	15	full	tumor protein p53 binding protein 1	onc
TYMS		31	28	26	24	22	full	THYMIDYLATE SYNTHETASE	fol
UGT1A1	UGT1A8	70	58	46	53	37	full	UDP glucuronosyltransferase 1 family, polypeptide A cluster	PAH
UGT1A8	UGT1A1	151	113	78	79	53	full	UDP GLUCURONOSYLTRANSFERASE 1 FAMILY, POLYPEPTIDE A8	PAH
UGT2B10		21	10	9	9	8	full	UDP GLUCURONOSYLTRANSFERASE 2 FAMILY, POLYPEPTIDE B10	nit/PAH
UGT2B11		18	10	9	10	9	full	UDP GLUCURONOSYLTRANSFERASE 2 FAMILY, POLYPEPTIDE B11	nit/PAH

Continued

TABLE 1

(*Continued*)

Gene symbol	Genetic locus overlaps	All HapTags: pairwise, MAF1%	Inf 0.6+, pairwise, MAF1%	Inf 0.6+, pairwise MAF5%	Inf 0.6+, multi, MAF1%	Inf 0.6+, multi, MAF5%	Array coverage: multi HapTags, MAF1%	Gene name	Target category[a]
UGT2B15		1	0	0	0	0	nine non-HapTag SNPs	UDP GLUCURONOSYLTRANSFERASE 2 FAMILY, POLYPEPTIDE B15	nit/PAH
UGT2B17		14	10	7	10	7	full	UDP GLUCURONOSYLTRANSFERASE 2 FAMILY, POLYPEPTIDE B17	nit/PAH
UGT2B28		8	4	4	4	4	full	UDP GLUCURONOSYLTRANSFERASE 2 FAMILY, POLYPEPTIDE B28	nit/PAH
UGT2B4		31	24	22	18	16	full	UDP GLUCURONOSYLTRANSFERASE 2 FAMILY, POLYPEPTIDE B4	nit/PAH
UGT2B7		21	17	15	13	11	full	UDP GLUCURONOSYLTRANSFERASE 2 FAMILY, POLYPEPTIDE B7	PAH
VCAM1		69	65	45	63	42	full	vascular cell adhesion molecule 1	adh/inf
VEGFA		45	45	36	43	33	full	VASCULAR ENDOTHELIAL GROWTH FACTOR A	onc
VEGFB		15	15	15	12	12	full	VASCULAR ENDOTHELIAL GROWTH FACTOR B	inf
VEGFC		73	71	63	54	45	full	VASCULAR ENDOTHELIAL GROWTH FACTOR C	inf
XIAP		25	18	18	18	18	full	BACULOVIRAL IAP REPEAT-CONTAINING 4	onc
XPA		34	31	25	26	20	full	XERODERMA PIGMENTOSUM, COMPLEMENTATION GROUP A	DNA
XPC		61	59	42	49	34	full	XERODERMA PIGMENTOSUM, COMPLEMENTATION GROUP C	DNA
XRCC1		47	46	30	42	27	full	X-RAY REPAIR COMPLEMENTING DEFECTIVE REPAIR IN CHINESE HAMSTER CELLS 1	DNA
XRCC4		134	120	94	94	68	full	X-ray repair complementing defective repair in Chinese hamster cells 4	DNA
	Sum:	**17,797**	**15,961**	**13,474**	**12,926**	**10,511**	**Count:**	**298**	

[a] adh, Adhesion molecules; DNA, repair of DNA damage; fol, folate transport and metabolism; inf, inflammatory signaling and processes or immune regulation; mut, mutagenic processes; nic, nicotine addiction and smoking behavior; nit, tobacco-specific nitrosamine (in particular, NNK) activation and detoxification; onc, oncogenesis; oxs, oxidative stress; str, steroid hormone metabolism and signaling; tox, other toxin or toxicity; tum, risk for lung cancer or related tumors.

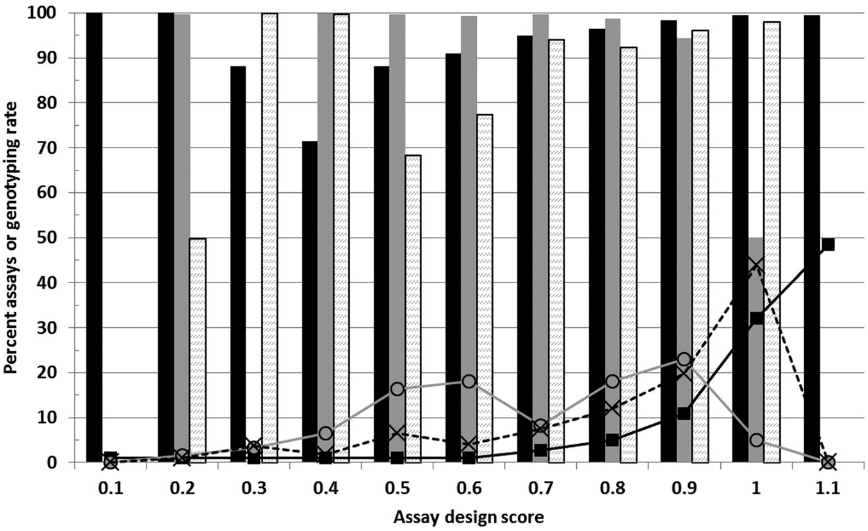

FIGURE 1

Distribution of assay conversion rates for SNPs in various design categories. Assays were assigned to Infinium design score bins equal to or less than the indicated values. The percent of all assays in a bin that successfully generated genotypes (unambiguous SNP allele calls in at least 95% of DNA samples) is plotted for Infinium-eligible SNPs in the Illumina database (black bars), mitochondrial DNA SNPs (gray bars), and SNPs uploaded as custom sequences (dashed bars). The number of SNPs in each bin, as a percent of total SNPs in each category, is plotted with square line markers for Infinium database SNPs, circles for mitochondrial, and X for custom sequences.

the array. All markers and their sequences, coordinates, and targeted genes are provided in Supplemental Table 2.

Genotyping assays were performed on 1873 DNA samples from lung-cancer patients and controls using LungCaGxE microarrays. Forty-seven samples had a SNP assay call rate <99.0%. If these samples are excluded, SNP assays with an Infinium design score of at least 0.6 produced unambiguous genotype calls in 99.03% of the attempted reactions (Fig. 1). Targeted functional SNPs with a design score <0.6 generated genotype calls in 84.96% of the attempted reactions; the average genotyping rate for SNPs with recognized rs numbers in the Illumina database was 99.09%, whereas the rate for SNPs submitted as custom sequences was 96.16% (design score >0.6 in both sets).

DISCUSSION

The advancement of array-based SNP genotyping technologies has led to genome-wide association studies (GWAS), in which genetic markers distributed evenly throughout the genome[17] (or covering predicted haplotypes throughout the genome[14]) are tested for statistically significant association with a phenotype. Arrays offer advantages for GWAS over current deep-sequencing methods, including lower cost, faster assay turnaround and sample throughput, and easier data processing. However, the success of proxy markers depends on linkage to causal but unmeasured genetic variants, and even the highest capacity arrays of over 5 million SNPs may not cover rare variants or diverse populations well. Whole-genome or exome sequencing directly detects causal variants and polymorphism types beyond bi-allelic single nucleotides and does not rely on linked markers for statistical analysis. Whether deployed on SNP arrays or deep sequencing platforms, the primary concern for whole-genome assays is statistical power. Rare variants, multiple causes for the same phenotype, intergenic and multigene effects, and genetically mandated differential interactions between genes and environmental variables can all combine with multiple testing correction requirements to drive study population sizes to thousands or tens of thousands of subjects to adequately power GWAS.[18–21] Projects of this scale are an expensive proposition for arrays and would be extremely costly with deep sequencing even at the as-yet unattained goal of $1000/genome.

Comprehensive genotyping of targeted genes, by arrays or sequencing, takes advantage of high multiplex assay capacities to saturate targets with genetic markers. Hence, array data are less reliant on capturing a single, important linked marker while retaining rapid sample throughputs, and sequencing costs and efficiency are improved by focusing on a subset of genes rather than the whole genome. Depending on the size of the target panel and degree of saturation desired, custom arrays or sequencing can ease multiple testing penalties and reduce study population sizes necessary to achieve statistical power. Of course, the critical issue for this strategy is choosing which genes to assay. For the LungCaGxE panel, we chose genes involved in pathways relevant to responses to environmental stressors and saturated the resulting target panel with genetic markers as well as previously demonstrated functional and disease-associated variants.

The Illumina design score, whereas generally predictive of positive assay performance, underestimated the LungCaGxE genotype success rate achieved for Infinium-eligible tagSNPs and custom SNPs from the nuclear genome. The design scores were somewhat less positively predictive (i.e., further underestimated) of genotyping rates achieved for mitochondrial genome SNPs, which performed well over a wide range of design scores. The relatively high success rates for assays with design scores <0.6 indicate that for future targeted genotyping projects, failure to meet this overly stringent standard cutoff should not necessarily disqualify an assay if the specific SNP in question is important for the study goals.

In summary, the investigator tasked with designing a custom-targeted genotyping assay must balance several considerations. Given that the platform's multiplex capacity is often dictated by the project's budget, the investigator must select the marker types, thresholds for number of genes targeted, and MAF cutoffs that will provide the most efficient use of available assay resources. Several iterations of empirical design are usually needed to assess the impact of these parameters, and this process is aided by a streamlined bioinformatics workflow. Tagger Batch Assistant helps automate the retrieval of genetic coordinates for requested genes, managing genome build versions and providing an output format that easily interfaces with Tagger for marker prediction. The resulting Tagger files are then automatically processed to connect markers with the user's upstream gene annotations. We used this tool to optimize the LungCaGxE design through multiple versions, preserving sensitivity for marker MAFs as low as 1%, while reducing the number of SNPs required by using the Tagger multimarker haplotyping algorithm. This array enables rapid, cost-effective, and comprehensive genotyping of a panel of genes important for exploring genetic factors in lung cancer and the environmental influences that impact those factors.

ACKNOWLEDGMENTS

This work was funded by grant PA4100038714 from the Pennsylvania Department of Health and U.S. National Institutes of Health–National Institute of Environmental Health Sciences grant 5 P30 ES 013508-06 for the Center of Excellence in Environmental Toxicology. We thank David McGain and Kathakali Addya (Penn Molecular Profiling Facility) and Cecilia Kim (Children's Hospital of Philadelphia Center for Applied Genomics) for technical assistance.

DISCLOSURE

The authors have no associations or sources of financial support that pose a conflict of interest for conducting or interpreting the work presented in this manuscript.

REFERENCES

1. American Cancer Society. *Cancer Facts & Figures 2013*. Atlanta, GA, USA: American Cancer Society, 2013 (http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-036845.pdf).
2. Cassidy A, Duffy SW, Myles JP, Liloglou T, Field JK. Lung cancer risk prediction: a tool for early detection. *Int J Cancer* 2007;120:1–6.
3. Ihsan R, Chauhan PS, Mishra AK, et al. Multiple analytical approaches reveal distinct gene-environment interactions in smokers and non-smokers in lung cancer. *PLoS One* 2011;6: e29431.
4. Thomas L, Doyle LA, Edelman MJ. Lung cancer in women: emerging differences in epidemiology, biology, and therapy. *Chest* 2005;128:370–381.
5. Braithwaite KL, Rabbitts PH. Multi-step evolution of lung cancer. *Semin Cancer Biol* 1999;9:255–265.
6. Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. *J Natl Cancer Inst* 2003;95:470–478.
7. Bilello KS, Murin S, Matthay RA. Epidemiology, etiology, and prevention of lung cancer. *Clin Chest Med* 2002;23:1–25.
8. Liu G, Zhou W, Christiani DC. Molecular epidemiology of non-small cell lung cancer. *Semin Respir Crit Care Med* 2005;26: 265–272.
9. Taioli E. Gene-environment interaction in tobacco-related cancers. *Carcinogenesis* 2008;29:1467–1474.
10. Gustafson AM, Soldi R, Anderlind C, et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. *Sci Transl Med* 2010;2:26ra25.
11. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;4:44–57.
12. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009;37:1–13.
13. De Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. *Nat Genet* 2005;37:1217–1223.
14. Peiffer DA, Le JM, Steemers FJ, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* 2006;16:1136–1148.
15. Goode EL, Fridley BL, Sun Z, et al. Comparison of tagging single-nucleotide polymorphism methods in association analyses. *BMC Proc* 2007;1(Suppl 1):S6.
16. Nam MH, Won HH, Lee KA, Kim JW. Effectiveness of in silico tagSNP selection methods: virtual analysis of the genotypes of pharmacogenetic genes. *Pharmacogenomics* 2007;8:1347–1357.
17. Matsuzaki H, Dong S, Loi H, et al. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* 2004;1: 109–111.
18. Becker T, Herold C, Meesters C, Mattheisen M, Baur MP. Significance levels in genome-wide interaction analysis (GWIA). *Ann Hum Genet* 2011;75:29–35.
19. Park JH, Wacholder S, Gail MH, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet* 2010;42:570–575.
20. Sale MM, Mychaleckyj JC, Chen WM. Planning and executing a genome wide association study (GWAS). *Methods Mol Biol* 2009;590:403–418.
21. Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* 2009;5:e1000477.