Molecular Therapy
Oncolytics
Original Article

AMERICAN SOCIETY of GENE & CELL THERAPY

Programmable Nuclease-Based Integration into Novel Extragenic Genomic Safe Harbor Identified from Korean Population-Based CNV Analysis

Eun-Seo Lee,[1,2,11] Sanghoon Moon,[3,11] Kwaku Dad Abu-Bonsrah,[4] Yun Kyoung Kim,[3] Mi Yeong Hwang,[3] Young Jin Kim,[3] Seokjoong Kim,[5] Nathaniel S. Hwang,[2,6,7] Hyongbum Henry Kim,[1,8,9,10] and Bong-Jo Kim[3]

[1]Department of Pharmacology, Yonsei University College of Medicine, Seoul 03372, Republic of Korea; [2]School of Chemical and Biological Engineering, Seoul National University, Seoul 08826, Republic of Korea; [3]Division of Genome Research, Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do 28159, Korea; [4]Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, VIC 3052, Australia; [5]ToolGen, Seoul 08501, Republic of Korea; [6]Interdisciplinary Program in Bioengineering, Seoul National University, Seoul 08826, Republic of Korea; [7]BioMax Institute of Seoul National University, Seoul 08826, Republic of Korea; [8]Brain Korea 21 Plus Project for Medical Sciences, Yonsei University College of Medicine, Seoul 03372, Republic of Korea; [9]Center for Nanomedicine, Institute of Basic Science (IBS), Seoul 03772, Republic of Korea; [10]Severance Biomedical Science Institute, Yonsei University College of Medicine, Seoul 03372, Republic of Korea

Here, we found two genomic safe harbor (GSH) candidates from chromosomes 3 and 8, based on large-scale population-based cohort data from 4,694 Koreans by CNV analysis. Furthermore, estimated genotype of these CNVRs was validated by quantitative real-time PCR, and epidemiological data examined no significant genetic association between diseases or traits and two CNVRs. After screening the GSH candidates by *in silico* approaches, we designed TALEN pairs to integrate EGFP expression cassette into human cell lines in order to confirm the functionality of GSH candidates in an *in vitro* setting. As a result, transgene insertion into one of the two loci using TALEN showed robust transgene expression comparable to that with an AAVS1 site without significantly perturbing neighboring genes. Changing the promoter or cell type did not noticeably disturb this trend. Thus, we could validate two CNVRs as a site for effective and safe transgene insertion in human cells.

INTRODUCTION

One of the critical needs in the biomedical field is the ability to stably insert functional transgenes and other genetic elements into the human genome without disrupting genes or perturbing their transcription, which can potentially alter the biological properties of host cells. Several diseases have been successfully treated with stable insertion of therapeutic genes, such as Leber's congenital amaurosis,[1] adrenoleukodystrophy,[2] and Parkinson's disease.[3] Furthermore, this stable gene insertion can facilitate determination of transgene functions, labeling cells for tracking and lineage analysis through reporter gene insertion, and production of specific proteins from human cells.

In the meantime, random integration of transgenes may cause insertional mutagenesis, which can possibly alter expression levels of neighbor genes. The most common approach of integrating transgenes in human cells is to use retroviral vector, which inserts transgenes into the human genome in a semi-random manner with a preference toward the vicinity of transcriptionally active genes.[4–6] This uncontrolled insertional mutagenesis can cause cancer such as leukemia[7] and lymphoma[8] by perturbing proto-oncogenes or tumor suppressors. Furthermore, when the transgenes are integrated into random regions of human chromosomes, the expression of these transgenes can be silenced or unpredictable depending on the integration site.

Instead of viral vectors, engineered nucleases such as zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and the CRISPR-Cas system have become more prospective approaches for the purpose of safety. These engineered nucleases induce site-specific double-stranded breaks into targeted sites and go through a highly precise genomic editing by the mechanism of homology-directed repair (HDR) in the presence of single-stranded oligodeoxynucleotide (ssODN) or a donor DNA.[9]

Despite the advance in these engineered nucleases, there are only a few up-to-date identified and validated genomic safe harbors (GSHs). GSHs are intra- or extragenic regions that can support predictable transgene expression while minimizing neighboring gene perturbation.[10] The AAVS1 site on chromosome 19 is the most popular GSH, due to its ability to support transgene expression in

Received 18 December 2018; accepted 11 July 2019;
https://doi.org/10.1016/j.omto.2019.07.001.

[11]These authors contributed equally to this work.

Correspondence: Hyongbum Henry Kim, Department of Pharmacology, Yonsei University College of Medicine, Seoul 03372, Republic of Korea.
E-mail: hkim1@yuhs.ac
Correspondence: Bong-Jo Kim, Division of Genome Research, Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do 28159, Korea.
E-mail: kbj6181@cdc.go.kr

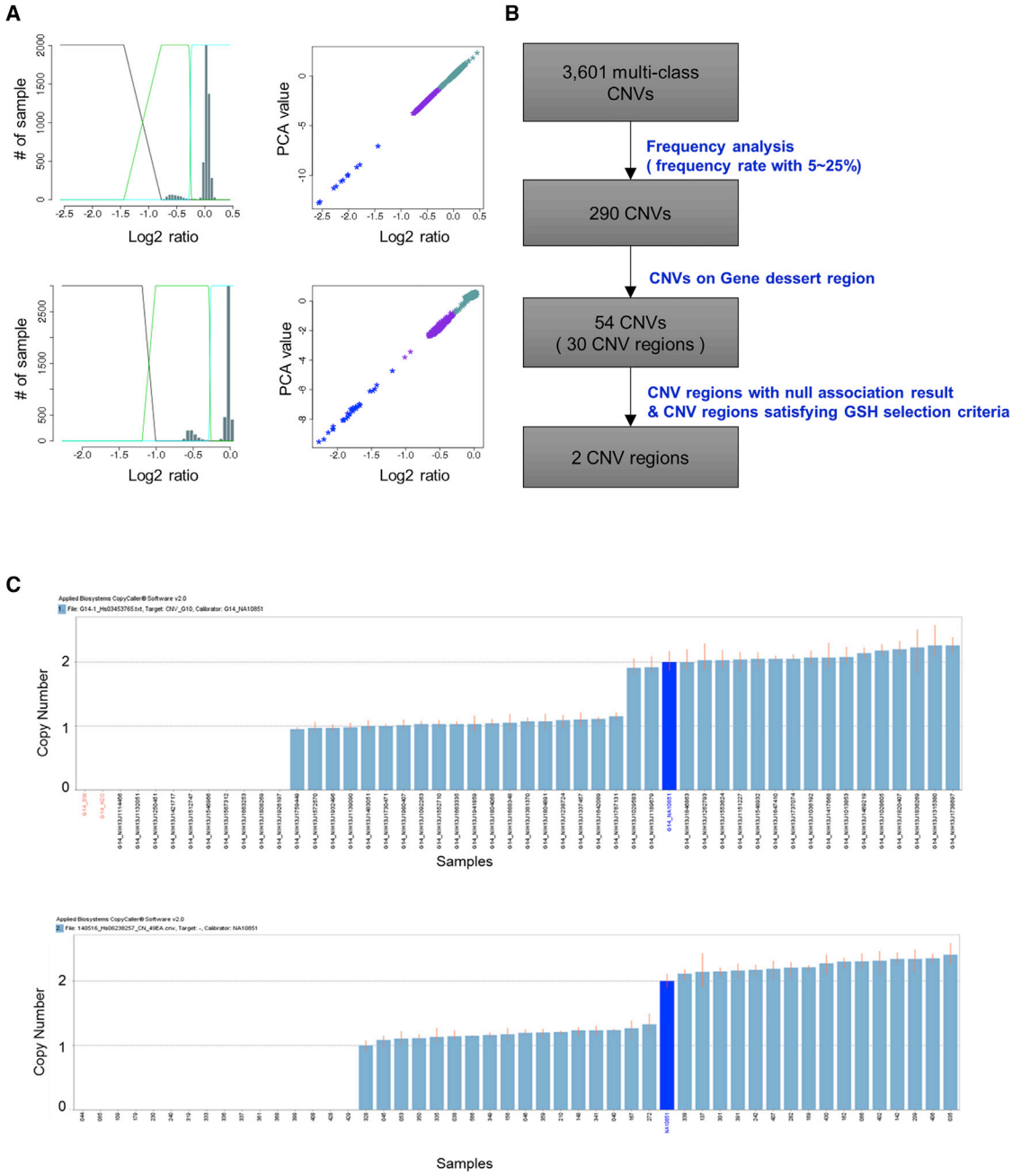

Figure 1. Estimated CNV Classes of Selected Regions and Validation of CNV Genotype of Each Individual

(A) The flowchart shows how various CNVs were screened to select two final GSH candidate regions. (B) Histograms and cluster plots illustrates two candidate regions of CNVR7 and CNVR22 generated by CNV tools. Like $\log_2$ ratio plots, histograms and cluster plots show that CNV genotypes of these regions were clearly separated in to three groups (0 copies, 1 copy, and 2 copies). (C) Quantitative real-time PCR data show validation results on CNVR7 and CNVR22. We conducted a validation experiment for three CNV states. Samples in each state were randomly selected. Higher bars, lower bars, and no bar mean normal copy (2 copies), heterozygous deletion (1 copy), and homozygous deletion (0 copies), respectively. Blue bar represents CNV genotype of the reference sample (NA10851) used by comparative genomic hybridization array (aCGH).

multiple cell types,[11,12] yet at the same time, it was known that the AAVS1 locus could be silenced by the mechanism including DNA methylations.[13]

A copy-number variation (CNV) is an insertion or deletion of DNA segment and is relatively common and widespread in the human genome.[14] There are different gene copy numbers in a particular

Table 1. List of 30 CNV Regions on Gene Desert Regions

CNVRID	Chromosome	Start	Stop	Frequency		
				Cluster 1	Cluster 2	Cluster 3
CNVR1	2	41091503	41105475	92.35	7.52	0.13
	2	41091926	41095029			
	2	41091926	41110876			
CNVR2	2	107383090	107385789	83.66	15.76	0.58
CNVR3	2	107633827	107644568	80.17	18.79	1.04
	2	107640592	107644729			
	2	107641774	107644004			
CNVR4	2	130094781	130097703	95.01	4.88	0.11
CNVR5	2	194397392	194400503	80.14	19.86	–
	2	194397392	194403581			
	2	194399479	194403581			
	2	194401599	194403581			
CNVR6	2	195688911	195690761	84.17	15.02	0.81
CNVR7*	3	82951465	82955620	90.82	8.95	0.23
CNVR8	3	112723878	112730315	85.28	14.23	0.49
CNVR9	4	25899244	25903905	81.85	17.53	0.62
	4	25900330	25903009			
CNVR10	4	52355328	52358561	86.39	13.61	–
	4	52355328	52360772			
	4	52355328	52378210			
	4	52356185	52360772			
CNVR11	4	61182975	61184206	90.84	8.97	0.19
CNVR12	4	61668290	61696232	88.13	11.53	0.34
CNVR13	4	64376477	64399338	75.48	23.11	1.41
	4	64376893	64389954			
	4	64376893	64402277			
	4	64381512	64384210			
	4	64386390	64392883			
CNVR14	4	138310481	138323629	76.01	22.09	1.90
	4	138311724	138316056			
	4	138315738	138319841			
CNVR15	5	57361273	57365938	75.16	24.84	–
CNVR16	5	97961273	97963268	93.52	6.35	0.13
CNVR17	6	14853578	14855073	93.69	6.31	–
CNVR18	6	95345973	95348593	92.99	6.84	0.17
CNVR19	7	86077403	86080393	86.13	13.23	0.64
CNVR20	7	144547892	144552203	93.95	5.99	0.06
	7	144549963	144552203			
CNVR21	8	2626786	2640303	89.84	9.91	0.26
CNVR22*	8	135127147	135140206	83.85	15.53	0.62
CNVR23	8	138195052	138196533	89.18	10.57	0.26
CNVR24	8	142926455	142932099	94.91	5.0	0.04
CNVR25	9	81218373	81223517	86.39	12.95	0.66
	9	81219657	81222228			
	9	81220275	81223087			
	9	81221187	81222270			
CNVR26	10	58571948	58610908	91.63	8.24	0.13
CNVR27	10	91988373	91992382	92.91	6.99	0.11
CNVR28	13	49967437	49970106	76.48	23.52	–
	13	49967836	49969229			
CNVR29	13	103074226	103077048	92.16	7.71	0.13
	13	103074677	103076638			
CNVR30	18	62928265	62929768	94.95	4.92	0.13

Of these regions, the two CNV regions of CNVR7 and CNVR22, shown with an asterisk, were selected for the GSH experiment. Frequency of each cluster has been rounded off to the nearest hundredth.

region among healthy individuals. We hypothesized that transgene insertion into the CNV region (CNVR), which has less association with genetic diseases, might not lead to any abnormal health problems. Accordingly, we performed the large-scale cohort studies through Korean CNV analysis to screen potential GSH candidates based on essential criteria.[10]

We selected two possible chromosomal sites, chromosomes 3 and 8, further away from both gene-rich regions and genes implicated in cancer and microRNA (miRNA). Once these chromosomal sites satisfied the criteria for GSH, they were further investigated to rule out any disease correlation. We were able to achieve efficient site-specific integration and to measure the neighboring gene perturbation near the site of integration. Consequently, we proved that our CNVRs could derive robust transgene expression without significantly altering multiple neighboring genes.

RESULTS

Selected CNVRs for GSH Site

Through a frequency analysis of CNV state, 290 CNVRs with 5%–25% frequency were selected from 3,601 CNVRs (Figure 1A; Figure S1A). Among those, 30 CNVRs had no neighboring genes within 300 kb upstream and downstream of a CNVR (Table 1). Subsequently, two CNV regions with much lower disease association through statistical analysis assessing the disease association, CNVR7 (chr3:82951465–82955620, hg18) and CNVR22 (chr8:135127147–135140206, hg18), were selected as final candidates (Table 2; Figure S1B).

Table 2 describes two selected regions. CNV genotypes of two regions were composed of homozygous deletion (0 copies), heterozygous deletion (1 copy), and normal copy (2 copies). Frequency of copy number deletion (cluster 1 + cluster 2) was 9.18% and 16.15% for CNVR 7 and CNVR22 (Table 2). Moreover, these regions overlapped

Table 2. Frequency Rate of Two Selected CNV Regions

CNVR ID	Chromosome	Locus	Start	End	Length (bp)	No. of Probes in the CNV Region	CNV Type	Frequency (n = 4,694)			DGV
								Normal (2 copies)	Heterozygous Deletion (1 copy)	Homozygous Deletion (0 copies)	
CNVR7	3	3p12.2	82951465	82955620	4,155	34	deletion	90.82% (4263)	8.95% (420)	0.23% (11)	overlapped
CNVR22	8	8q24.22	135127147	135140206	13,060	38	deletion	83.85% (3936)	15.53% (729)	0.62% (29)	overlapped

Major CNV state is normal copy, followed by heterozygous deletion and homozygous deletion. Frequency of homozygous deletion is less than 1% in both regions.

with previously reported CNV regions found in the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home). A length of CNV region at chromosome 3 was about 4.1 kb, and 34 consecutive probes were included in this region; the length of CNV region at chromosome 8 was 13 kb, and 38 probes were included (Table 2).

Validation of Estimated CNV Genotype on Two Selected Regions by Quantitative Real-Time PCR

Figures 1B and 1C represent estimated CNV classes and validated CNV genotypes. We estimated that there were three copy-number classes in two selected regions. Moreover, to evaluate whether estimated CNV genotype of each individual is concordant with real CNV genotype, we randomly selected 49 and 47 samples from CNVR7 and CNVR22, respectively. To examine overall accuracy, positive predictive value (PPV) was used as the measurement standard of accuracy. PPV was defined as the proportion of true-positive numbers to number of positive calls.

From quantitative real-time PCR experiment results, we confirmed that two regions consisted of three copy-number classes and the CNV genotype of each individual was perfectly matched to those of our estimation (Figure 1C). PPV of each candidate region was 1.

Disease-Association Results of Two Candidate Regions

Tables 3 and 4 are association analyses of CNVR7 and CNVR22, respectively. In the case of CNVR7, most diseases such as type 2 diabetes (T2D), hypertension, osteoporosis, obesity, dyslipidemia, and metabolic syndrome were not significantly associated with this region, except for high-density lipoprotein (HDL). The p value of HDL-CNV association analysis was $p < 0.05$ ($p = 0.037$) (Table 3). Because the sample size of this case-control study might cause spurious results, we also checked statistical significance of association by conducting linear regression analysis. The results from linear regression analysis exhibited conflicting results ($p = 0.178$) compared to those from logistic regression. From this, we assumed that this discrepancy indicates no CNV association with HDL trait. In the case of CNVR22, we found that there was no statistical significance in association results (Table 4). Tables S1–S28 illustrate disease-association analysis of 28 CNVRs.

EGFP Cassette Integration and Expression into AAVS1, CNVR7, and CNVR22

We designed TALEN pairs to target AAVS1, CNVR22, and CNVR7 sites in human somatic cell lines and their corresponding EGFP expression cassettes, which were driven by the CMV early enhancer or chicken beta-actin (CAG) and viral origin SFFV (spleen focus-forming virus) promoter terminating at the poly(A) site from the bovine growth hormone gene (Figure 2; Figure S2B). To promote efficient homologous recombination at the required locus of AAVS1, CNVR7, and CNVR22, each of the targeting donor cassettes was constructed with homology arms of 800 bp (left arm) and 800 bp (right arm). The appropriate combination of TALENs and the EGFP expression cassette was transiently transfected into human K562 cells and Huh 7.5 cells by electroporation. We derived single clones from $EGFP^+$ sorted cells and performed site-specific integration PCR analysis to confirm transgene integration at the target site by homology-directed repair (Figure S2C). From K562 cells, we selected a total of 252, 184, and 223 clones for AAVS1, CNVR7, and CNVR22, respectively, and sorted $EGFP^+$ clones (Figures S2A and S2B). A similar degree of EGFP-expressing populations among three sites may validate the use of two extragenic GSH candidate regions. Then, target-integrated clones from $EGFP^+$ clones were represented by 78.6%, 73.1%, and 79% for AAVS1, CNVR7, and CNVR22 (Figure 3A). As reported by Lombardo et al.,[15] integration efficiency (percentage of $EGFP^+$ cells) is unaffected by the target site, and mean fluorescence intensity (MFI; a measure of the average expression per cell) of EGFP depends on both the promoter and the target locus. We also recorded a similar outcome (more than 70% of confirmed transgene integrated clones for each construct)[15] in that the MFI of EGFP showed no significance between AAVS1 and CNVR22 sites but outperforming CNVR7, with $p = 0.0256$ (Figure 3B). Nonetheless, the integration efficiency was dependent on the TALEN activity (data not shown).

To validate that such findings are not strictly defined to one cell type, we tried a different cell line, Huh 7.5, that originated from a different human tissue.[10] Site-specific integration PCR analysis was performed on single cell-derived clones, and a similar finding was reported as that seen in K562 cells (Figure S3B). Out of the $EGFP^+$ sorted clones, 11/23 for AAVS1, 4/18 for CNVR7, and 6/20 for CNVR22 were $EGFP^+$-targeted integrations representing 47.8%, 17.8%, and 26.1% (Figures 3C; Figure S3A), respectively, as analyzed by flow cytometry. The MFI of $EGFP^+$ cells showed significance between AAVS1 and CNVR7 sites ($p = 0.0032$) and between CNVR7 and CNVR22 sites ($p = 0.0092$) (Figure 3C).

Extragenic Integration into CNVR7 and CNVR22 Did Not Regulate Nearby Genes as It Did into AAVS1

The AAVS1 site has already been reported as a possible safe harbor site, where integration into the AAVS1 site leads to stable expression

Table 3. Disease-Association Results of CNVR7 at Chromosome 3

Disease Criteria	Trait	Type	Beta	SE	p Value
Diabetes	GLU	CC	0.3723	0.190	0.051
Hypertension	HTN	CC	0.04717	0.116	0.685
Osteoporosis	AS1_DT_cc	CC	0.1335	0.191	0.484
	AS1_MT_cc	CC	0.1018	0.140	0.469
Obesity	BMI	QT	−0.02699	0.156	0.863
Dyslipidemia	HDL	CC	−0.2041	0.098	0.037*
	HDL_QT	QT	−0.02704	0.020	0.178
	TCHL	CC	0.2272	0.151	0.133
	TG	CC	−0.05246	0.129	0.684
	LDL	CC	0.2654	0.179	0.139
Metabolic syndrome	MS_cc	CC	−0.1092	0.106	0.303
Tumor	–	–	–	–	–
Respiratory disease	AS1_BrDs	CC	−0.2258	0.122	0.065
Joint disease	AS1_DgnArth	CC	0.1001	0.151	0.507
	AS1_RhmArth	CC	−0.2885	0.189	0.127
Insomnia	AS1_Insm	CC	−0.143	0.126	0.255
Clinical test (blood and urea)	AS1_BCTRIA	CC	0.3731	0.233	0.109
	AS1_CRYSTAL1	CC	−0.09184	0.140	0.513
	AS1_CRYSTAL2	CC	−0.1044	0.168	0.535
	AS1_CRYSTAL3	CC	−1.4017	0.633	0.027*
	AS1_CRYSTAL4	CC	0.1265	0.252	0.616
	AS1_CRYSTAL5	CC	−0.5284	0.736	0.470
	AS1_U_OTHR	QT	0.002803	0.006	0.632
	AS1_VB12	QT	30.2	92.440	0.744
	AS1_FOLATE	QT	1.622	2.363	0.494
	AS1_VDRL	CC	15.76	2,982.630	0.996
	AS1_FREET4	QT	0.04362	0.056	0.436
	AS1_TSH	QT	0.2098	0.254	0.409
	AS1_CD	QT	0.5005	0.434	0.252
	AS1_PB	QT	0.4125	0.618	0.506
	AS1_AL	QT	−0.2763	0.244	0.261
Body metrics	HEIGHT	QT	−0.2409	0.435	0.580
	WEIGHT	QT	−0.2539	0.506	0.616
Lung function test	AS1_SP1_3	QT	0.3502	0.756	0.643
	AS1_SP2_3	QT	0.1732	0.906	0.848
	AS1_SP3_1	QT	−0.08605	0.092	0.351
Electrocardiogram	AS1_EKG	CC	0.1157	0.115	0.314
Chest X-ray	AS1_CH0	CC	−0.07924	0.105	0.449
Gender	SEX	QT	−0.000376	0.024	0.988
Age	AGE	QT	0.4554	0.440	0.301
Wearing glasses	AS1_Glasses	CC	−0.033432	0.098	0.733
Hearing aid	AS1_Acst	CC	−0.4761	0.396	0.229
Been in accidents	AS1_AccFq	QT	0.04335	0.101	0.668
Tooth problem	AS1_Tooth	QT	−0.06614	0.038	0.084
Medical history	AS1_PdHn	CC	−0.4883	0.521	0.348
	AS1_PdUt	CC	−0.6173	0.469	0.188
	AS1_PdGt	CC	−0.2853	0.169	0.091
Current medical diagnosis and treatment	AS1_PdIm	CC	−0.3394	0.518	0.512
	AS1_TrtAr	CC	−0.1249	0.256	0.626
	AS1_TrtGt	CC	15	3,563.75	0.997

A p value of each trait has been rounded off the numbers to the nearest thousandth. *p < 0.05. QT, quantitative trait; CC, case control.

with no upregulation of nearby genes.[15] We next investigated whether extragenic integration into CNVR7 and CNVR22 in K562 cells would lead to a stable and reliable expression with no neighboring gene perturbation beyond 300 kb up- and downstream of the target locus. We analyzed 9 genes for AAVS1 found within 300 kb upstream and downstream of the integration site by quantitative real-time PCR. 5 out of 9 genes in AAVS1-targeted cells demonstrated a significant downregulation, while PPPIR12C at the targeted integration point exhibited the steepest downregulation (Figure 4A).

In order to satisfy the criteria for safe harbor (Figure S4A), CNVR7 and CNVR22 were found to not contain any nearest gene within 300 kb up- and downstream of the integration site. For example, as to CNVR7 and CNVR22, the nearest genes to the target locus are GBE1 and ZFAT, respectively, whose distances from the target locus are 1.05 Mb and 426 kb (Figure S4B). After integrating a CAG-promoter-driven EGFP cassette into CNVR7, we observed upregulation of CADM2 and substantial downregulation of ROBO1 (Figure 4B). When we additionally analyzed 6 single-cell-derived clones with molecularly confirmed targeted integration in this locus, we could acquire similar outcome (data not shown). This demonstrates that transcriptional upregulation of the locus is independent of the extent of EGFP expression, as mentioned earlier. Performing the same analysis on single-cell-derived clones with molecularly confirmed targeted integration in CNVR22 resulted in no significant neighbor gene dysregulation (Figure 4C). Therefore, a potency in stable and robust transgene expression followed by specific target integration in our safe-harbor candidates of CNVR7 and CNVR22 would be comparable to AAVS1.

To extend the generality of these findings, we assayed the impact of equivalent CAG-promoter-driven EGFP cassette integration on the three target loci in a different cell line, Huh 7.5 (Figure 4D). We observed significant perturbation in 4 out of 9 genes in AAVS1-targeted cells and 1 out of 5 genes in CNVR7-targeted cells. In CNVR22-targeted cells, single-cell-derived clones were not perturbed as much as in other loci, but we observed 13.5-fold downregulation of the flanking gene, TG. Again, even though we could observe one significant perturbation in both CNVR7 and CNVR22, it did not invoke the perturbations as much as seen in AAVS1, and this further supports that our GSH candidates are comparable to AAVS1.

Table 4. Disease-Association Results of CNVR22 at Chromosome 8

Disease Criteria	Trait	Type	Beta	SE	p Value
Diabetes	glu0	CC	−0.08322	0.185	0.653
Hypertension	HTN	CC	0.01781	0.097	0.854
Osteoporosis	AS1_DT_cc	CC	−0.01528	0.169	0.928
	AS1_MT_cc	CC	0.06409	0.118	0.587
Obesity	bmi	QT	0.08126	0.129	0.529
Dyslipidemia	hdl	CC	−0.1087	0.081	0.178
	tchl	CC	−0.03003	0.134	0.823
	tg	CC	−0.02894	0.103	0.779
	ldl	CC	0.1186	0.148	0.423
Metabolic syndrome	MS_cc	CC	0.08018	0.085	0.348
Tumor	–	–	–	–	–
Respiratory disease	AS1_BrDs	CC	0.06889	0.108	0.525
Joint disease	AS1_DgnArth	CC	−0.04364	0.119	0.714
	AS1_RhmArth	CC	−0.02562	0.173	0.882
Insomnia	AS1_Insm	CC	−0.09224	0.106	0.384
Clinical test (blood and urea)	AS1_BCTRIA	CC	−0.2135	0.155	0.168
	AS1_CRYSTAL1	CC	−0.05077	0.118	0.667
	AS1_CRYSTAL2	CC	−0.1755	0.138	0.202
	AS1_CRYSTAL3	CC	0.387	1.061	0.715
	AS1_CRYSTAL4	CC	0.1952	0.211	0.354
	AS1_CRYSTAL5	CC	0.7935	1.042	0.446
	AS1_U_OTHR	QT	0.002045	0.005	0.673
	AS1_VB12	QT	67.04	71.98	0.353
	AS1_FOLATE	QT	1.43	1.844	0.439
	AS1_VDRL	CC	15.84	2,242	0.994
	AS1_FREET4	QT	−0.01144	0.044	0.794
	AS1_TSH	QT	−0.1704	0.198	0.390
	AS1_CD	QT	−0.1904	0.334	0.569
	AS1_PB	QT	0.2346	0.473	0.621
	AS1_AL	QT	−0.01982	0.188	0.916
Body metrics	height	QT	−0.02356	0.359	0.948
	weight	QT	0.142	0.418	0.734
Lung function test	AS1_SP1_3	QT	0.5637	0.626	0.368
	AS1_SP2_3	QT	0.04428	0.75	0.953
	AS1_SP3_1	QT	0.07714	0.076	0.313
Electrocardiogram	AS1_EKG	CC	0.04547	0.097	0.639
Chest X-ray	AS1_CH0	CC	0.12262	0.089	0.170
Gender	sex	QT	0.002885	0.02	0.886
Age	age	QT	−0.2904	0.364	0.425
Wearing glasses	AS1_Glasses	CC	−0.03577	0.081	0.658
Hearing aid	AS1_Acst	CC	0.07041	0.411	0.864
Been in accidents	AS1_AccFq	QT	−0.08382	0.082	0.305
Tooth problem	AS1_Tooth	QT	0.01922	0.032	0.542
Medical history	AS1_PdHn	CC	−0.2133	0.497	0.668
	AS1_PdUt	CC	−0.8534	0.401	0.033*
	AS1_PdGt	CC	0.2599	0.168	0.123
	AS1_PdIm	CC	0.5432	0.609	0.372
Current medical diagnosis and treatment	AS1_TrtAr	CC	0.1166	0.229	0.610
	AS1_TrtGt	CC	15.12	2,783	0.996

A p value of each trait has been rounded off the numbers to the nearest thousandth. *$p < 0.05$. QT, quantitative trait; CC, case control.

A Stronger Expression Promoter Did Not Affect the Perturbation

To test whether these findings were not promoter dependent, we designed an EGFP cassette with a stronger promoter, SFFV. Again, site-specific integration PCR analysis was performed on single-cell-derived clones in K562 cells (Figure S5B). Out of the $EGFP^+$ sorted clones, 29/45 for AAVS1, 9/23 for CNVR7, and 39/80 for CNVR22 were $EGFP^+$-targeted integrations representing 64.4%, 39.1%, and 48.8% (Figure 5A), respectively, as analyzed by flow cytometry. While the type of promoter did not affect target integration efficiency, MFI was higher in the SFFV promoter group than in the CAG promoter group (Figure S5A). The MFI of $EGFP^+$ cells showed significance between AAVS1 and CNVR7 sites ($p = 0.0079$) and between CNVR7 and CNVR22 sites ($p = 0.0464$).

Similar to the results shown in CAG-promoter-driven integration, AAVS1-targeted cells exhibited three significant gene perturbations (Figure 5B), whereas neither CNVR7 nor CNVR22 demonstrated significant gene dysregulation, except for the CNVR22 group's flanking gene, TG, which showed a similar degree of downregulation (13.1-fold). Therefore, changing the promoter type did not induce any detectable trend in neighboring gene perturbation in the three groups.

DISCUSSION

The integration of transgene into human chromosomes should be very careful, because it may cause unpredictable adverse effects, depending on the integration site. Hence, it is very important to check whether the integration leads to not only unpredictable effects on the cell but also undesirable outcomes for the human phenotype. Here, in order to find the GSH candidate region, we used real-world data based on already known CNVRs and concurrent disease statuses from cohort participants. Then, we performed the experimental analysis on transgene insertion into GSH candidates and validated their suitability by comparison to AAVS1.

We first proceeded with genomic approaches to find appropriate safe-harbor candidates, especially from the Korean population, as described in Figure S1A.[10,15] Depending on target cell type, transgene, or disease type, safe harbors may be sub-categorized into specific types, yet we aimed to identify universal GSHs for the general Korean population. Throughout the disease-association analysis in 4,694

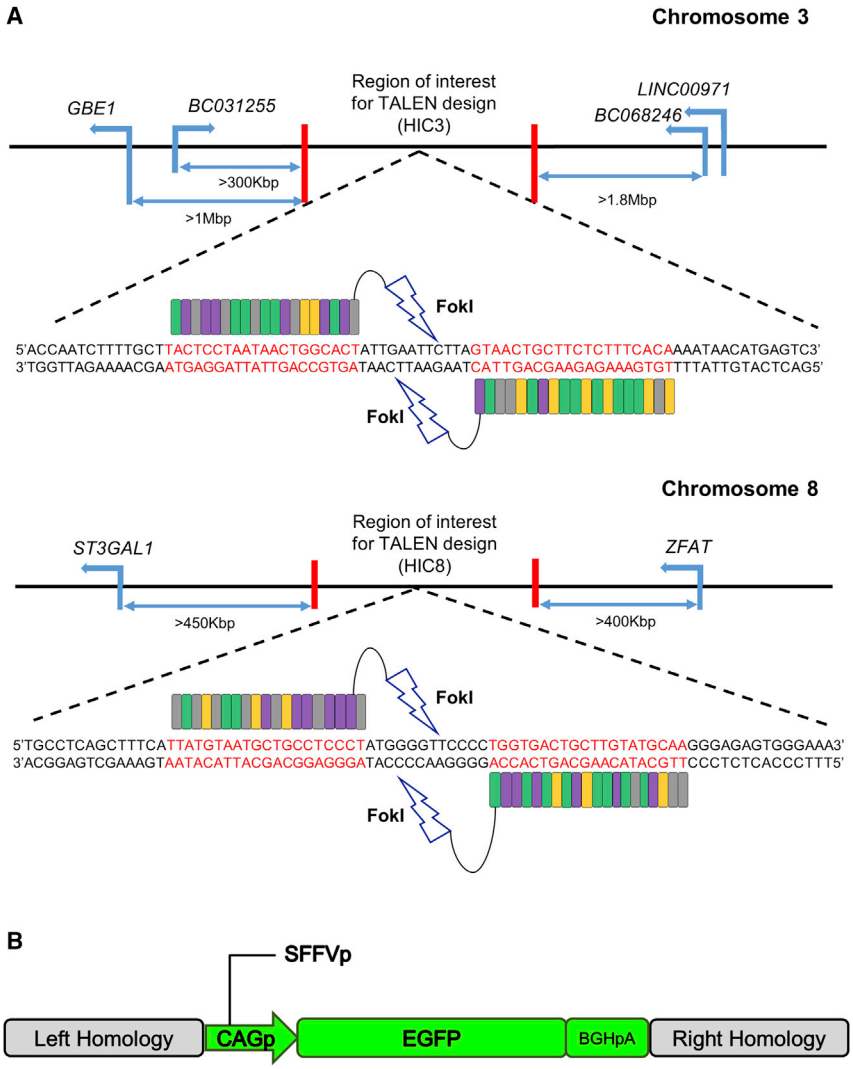

Figure 2. Site-Specific Integration and Transgene Expression of EGFP in AAVS1, Chromosome 3, and Chromosome 8 of K562 Cells

(A) Schematic illustrating the GSH candidate sites. Chromosome 8 is >450 kbp from the 3′ end of the *ST3GAL1* gene (gene nearby) and >400 kbp from the 5′ end of the *ZFAT* gene. On the other hand, chromosome 3 is >300 kbp from the 3′ end of BC031255 (a *Homo sapiens* cDNA clone), >1 Mbp from the 3′ end of GBE1, and >1.8 Mbp from the 5′ end of BC068246 (a *Homo sapiens* cDNA clone) and LINC00971 (long intergenic non-protein coding RNA 971). (B) The structure of the targeting construct (donor EGFP) contains the expression cassette (green box) flanked by the homology sequences (black boxes) to AAVS1, chromosome 3, and chromosome 8. BGHpA, poly(A) site from the bovine growth hormone gene.

Korean adults aged 40 to 69, we could select two safe-harbor candidates of CNVR7 and CNVR22. Since this *in silico* analysis preconfirmed that extragenic insertion into two selected regions barely influenced copy number change, we experimentally inserted the EGFP donor cassette by TALEN pairs into AAVS1, CNVR7, and CNVR22 to observe whether site-specific extragenic expression would incur any significant endogenous gene perturbation.

Our *in vitro* data demonstrate that both CNVR7 and CNVR22 are comparable to AAVS1. Especially, CNVR22 can be more feasibly utilized as a new GSH in human cells, in that neighboring gene expression beyond at least 426 kb on either side of the insertion site was not significantly expressed as maintaining adequate EGFP intensity. CNVR7 was enough to be categorized into GSH, yet less neighboring gene perturbation unwaveringly supports CNVR22 as a more suitable GSH candidate. After transfecting the EGFP donor cassette into CNVR7 and CNVR22 target loci, integrated clones were thoroughly analyzed to observe the neighbor gene expressions. Since CNVR7 and CNVR22 were chosen by safe-harbor criteria of excluding the genomic region within 300 kb of any miRNA and 500 kb of oncogenes, there was no genome within this range, unlike AAVS1, adjacent to the immediate vicinity within 100 kb. While the most common insertional oncogenesis comes from transactivation of neighboring oncogenes,[16] transgene insertion at CNVR7 and CNVR22 will not suffer from undesirable transcription caused by unwanted normal gene dysregulation, which tends to occur up to a distance of ~275 kb from the vector insertion site.[17] Although AAVS1 is known to resist neighboring gene perturbation due to the presence of chromatin insulator preventing enhancer-stimulated gene expression,[18] the AAVS1 locus was still exposed to uncontrolled gene dysregulation, as shown in our data. This further supports the suitability of CNVR7 and CNVR22 as comparable safe-harbor candidates to AAVS1.

When it comes to a broader application in clinical therapeutics, CNVR22 is well suited to the definition of universal GSH, owing to its appropriate expression independent of cell type and promoter. However, a reason for conspicuous TG downregulation upon altering the cell line and promoter is not clearly assumed, albeit TG-adjacent genes were not significantly perturbed. In domain-wide regulation spanning megabases, the activation or silencing of genes can be often accompanied by changing histone code or DNA methylation that can spread over considerable genomic distances, and this can disturb chromatin condensation.[13] Since TALEN binds to 5-methylated cytosine in its endogenous cognate target,[19] this could influence epigenetic modification, which may disturb endogenous TG expression. This phenomenon may also raise safety issues with regard to target locus containing transgene integration. In this sense, the CRISPR-Cas9

A

	GFP+ clones	Integrated Clones	Perc. (%)
AAVS1	61	48	78.6
CNVR7	41	30	73.1
CNVR22	58	46	79

C

	GFP+ clones	Integrated Clones	Perc. (%)
AAVS1	23	11	47.8
CNVR7	18	4	17.8
CNVR22	20	6	26.1

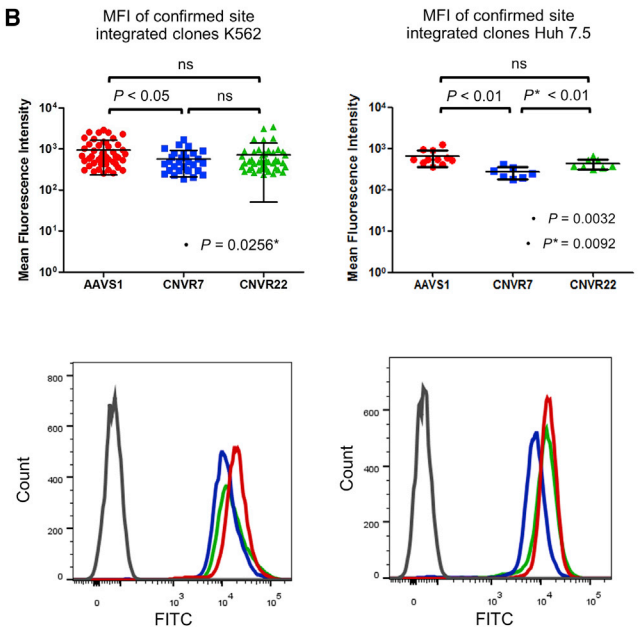

Figure 3. Site-Specific Integration and Transgene Expression of EGFP Cassette Driven by CAG Promoter

(A) Target integration percentage from EGFP$^+$ sorted clones in each target locus in K562 cells. (B) Scattered dot plot representing MFI from site-specific integration PCR confirmed single clones from K562-targeted cells after 3 weeks of transfecting EGFP$^+$ donor cassette and TALEN constructs (p = 0.0256, one-way ANOVA with Bonferroni's multiple comparison post-test). Graph indicates the mean ± SD. (C) For Huh 7.5 cells, target integration percentage from EGFP$^+$ sorted clones in each target locus and scattered dot plot representing MFI from site-specific integration PCR confirmed single clones from each target site after 3 weeks of transfecting EGFP$^+$ donor cassette and TALEN constructs. (p = 0.0032; *p = 0.0092, one-way ANOVA with Bonferroni's multiple comparison post-test). Graphs indicate the mean ± SEM.

system can be adapted to improve targeting efficiency and safety, and optimal design of the transgene cassette would be another alternative strategy to diminish endogenous transcription around the insertion site.

Throughout our GSH screening strategy, combining both genomic data with disease association in a general Korean population and experimental analysis on transgene insertion, we could discover the new GSH candidate of CNVR22 in the human genome and characterize putative universal GSHs preventing genotoxicity and achieving stable extragenic insertion near endogenous neighboring genes, which was less significantly disrupted than those of AAVS1. While such a rapid advance in genome editing field challenges us to constantly focus on seeking novel techniques for effective clinical studies, our findings on new GSHs specialized to Korean populations will broaden the horizon to search for prospective therapeutics by sustainable gene transfer.

MATERIALS AND METHODS

Samples and CNVRs

We first used the 3,601 CNVRs from the previous Korean CNV study using 4,694 individuals.[20] These were part of 8,842 Korea Association Resource (KARE) Project genome-wide association study (GWAS) subjects who satisfied quality control criteria, including exclusion of any kind of cancer samples (n = 101) from 10,038 subjects (Figure S1A).[21] Table 5 shows a summary of the 4,694 participants' characteristics. Moreover, the Supplemental Materials and Methods show CNV detection methods such as characteristics of genotyping array and CNV detection tools or parameters in detail. All procedures were in accordance with the ethical standards of the responsible committee on human experimentation (approved IRB number of Korea Centers for Disease Control and Prevention in Korea: 2016-02-20-T-A). Informed written consent was obtained from all participants.

Many of the copy number duplications are tandem, and this means that duplicated regions are located very near to each other. However, a significant number of copy number duplications are located far from the original locus in humans.[22] Moreover, these dispersed duplications appear randomly distributed among the genome.[23,24] A current CNV detection approach that we chose enabled us to discover copy number duplication but cannot determine the location of duplication. In this region of the initial set of CNVRs, we excluded copy number duplication regions.

CNVR Selection Criteria for GSH Candidate Regions

We postulated that a certain genome region could be a GSH candidate as long as deletion of a corresponding region does not lead to any abnormal health problems, including neoplasm. As some adverse phenotypes of genetic difference can be observed as the individuals become old, we analyzed the genome sequences from 4,964 of men and women who do not have a history of cancer. To fulfill the safety issue, we set the criteria for safe-harbor candidates to be more stringent than previously proposed[10] (Figure S1A). Then, we mainly considered well-genotyped common CNVRs (>5% frequency rate) to be 2- or 3-class CNV genotypes, with the distribution of minor CNV state ranging from 5% to 25%, because it is known that the common CNVs have a much lower effect on disease than rare CNVs.[24] Moreover, disease-association analysis with well-genotyped CNVRs is more reliable than that with poorly genotyped CNVRs. Regarding CNV with 3 classes, we calculated minor CNV state as the sum of two minor states. For example, if frequencies of CNV state with 3 classes were 85%, 12%, and 3%, we calculated minor CNV state as 15%.

To investigate CNVs on a gene-poor region, we used an annotation script of the PennCNV using refGene annotation of the NCBI Reference Sequence Database for human hg18 genome build.[25] We

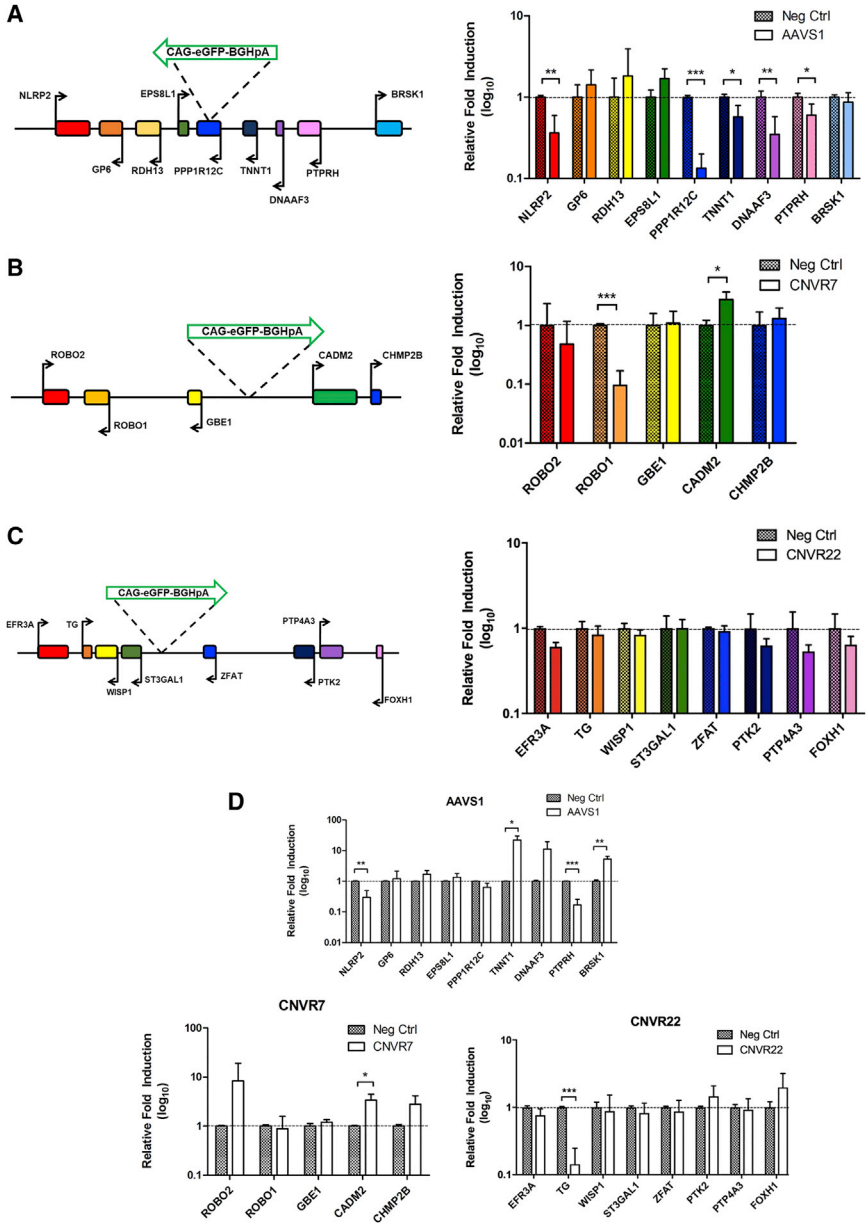

Figure 4. Endogenous Neighboring Gene Expressions Near the Integration Site

(A–C) Representation of genomic region on AAVS1 (A), CNVR7 (B), and CNVR22 (C) with their neighboring genes. Fold changes were analyzed by quantitative real-time PCR. EGFP$^+$ sorted K562 cells relative to mock-treated cells upon targeting the indicated cassettes into AAVS1, CNVR7, and CNVR22. For all genes, n = 7–8. Dashed line indicates the reference value in mock-treated cells. Graphs indicate the mean ± SEM. (D) Fold changes were analyzed by quantitative real-time PCR. EGFP$^+$ sorted Huh 7.5 cells relative to mock-treated cells upon targeting the indicated cassettes into AAVS1, CNVR7, and CNVR22. For all genes, n = 6. Dashed line indicates the reference value in mock-treated cells. Graphs indicate the mean ± SEM.

examined whether identified candidate CNV regions are correlated to disease-related genetic components or not. To do this, we selected CNVs with a neighboring distance of >300 kb from the 5′ end of any genes, including cancer-related genes, and a distance of >100 kb from any miRNA. We also selected CNVs outside of a gene transcription unit and ultra-conserved regions. Each CNV region was assessed for the existence of protein-coding genes and also non-coding RNA elements on the GENCODE database (human genome build HG38).

We also examined cancer-related elements near GSH candidate regions using the COSMIC genome browser of the COSMIC database (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/). Genomic locations for GSH candidate regions based on human genome assembly hg18 were converted to those based on hg19 by the liftOver tool from the UCSC genome browser. Cancer-related elements with a neighboring distance of >300 kb from each GSH candidate region were investigated.

To observe potential regulatory elements in regions around CNVs (±300 kb), we assessed common DNase I hypersensitivity regions (Broad ChromHMM, UW DNaseI DGF, and UW DNaseI HS) and transcription factor binding sites (Yale TFBS) among multiple cell types on the UCSC database.

Validation of Estimated CNV Genotype on Two Candidate Regions by Experiment

Estimated CNV genotypes had three copy-number classes, composed of homozygous deletion (0 copies), heterozygous deletion (1 copy), and normal copy (2 copies) (Figure 1B). To validate whether the estimated CNV genotype is true or not, we carried out quantitative real-time PCR using the TaqMan Copy Number Assay (Life Technologies, Foster City, CA, USA) according to the manufacturer's protocols. Two pre-designed TaqMan probes, Hs03453765_cn and Hs_06238257_cn, were used to validate the genotype of the two CNVs at chromosomes 3 and 8, respectively. All experiments were replicated three times to increase the validation accuracy. Moreover, validation samples were randomly selected from each CNV region, and genotypes of the reference DNA (NA10851) were observed together. CopyCaller v.2.0 (Life Technologies, Foster City, CA, USA) was used to analyze data generated by the TaqMan Copy Number Assay.

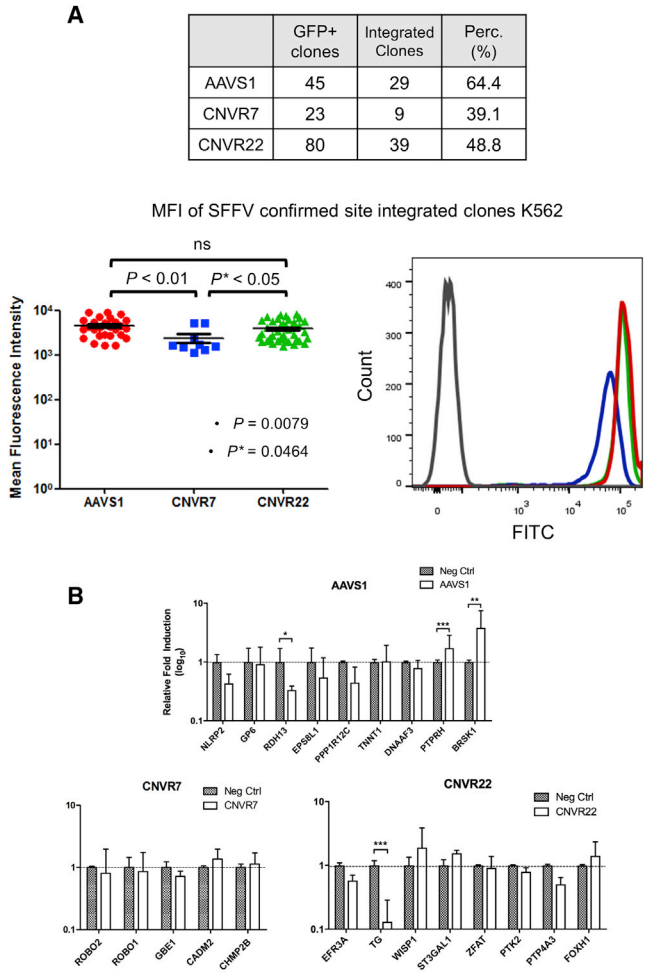

Figure 5. Site-Specific Integration and Transgene Expression of EGFP Cassette Driven by SFFV Promoter in K562 Cells

(A) Target integration percentage from EGFP$^+$ sorted clones in each target locus in K562 cells and scattered dot plot representing MFI from site-specific integration PCR confirmed single clones from each target site after 3 weeks of transfecting EGFP donor cassette and TALEN constructs (p = 0.0079; *p = 0.0464, one-way ANOVA with Bonferroni's multiple comparison post-test). Graph indicates the mean ± SD. (B) Fold changes were analyzed by quantitative real-time PCR. EGFP$^+$ sorted K562 cells relative to mock-treated cells upon targeting the indicated cassettes into AAVS1, CNVR7, and CNVR22. For all genes, n = 5–8. Dashed line indicates the reference value in mock-treated cells. Graphs indicate the mean ± SEM.

Statistical Analysis of Disease/Trait Association Using Epidemiology Data

Subsequently, statistical analyses were conducted to find a correlation between candidate CNV regions and diseases or traits. Using the epidemiological information of 4,694 individuals, we grouped datasets for each trait. In total, 23 diseases and traits, including T2D, hypertension, and obesity, were considered as well (Table 6). We did not conduct an association study for cancer, because 101 cancer patients and/or individuals with a medical history of cancer were already excluded. Logistic or linear regression analysis adjusting for gender and age as covariates was used to calculate statistical significance.

Table 5. Basic Characteristics of Samples Used in This Study

Traits	KARE CNV Study Total (N = 4,694)
Age (years)	54.0 ± 9.04
Male	2,210 (47.08%)
Female	2,484 (52.92%)
Height (cm)	159.5 ± 8.92
BMI (kg/m^2)	24.7 ± 3.20
SBP (mmHg)	121.3 ± 19.31
DBP (mmHg)	77.2 ± 11.91
Pulse rate (BPM)	64.2 ± 7.98
WHR	0.89 ± 0.07
ALT (IU/L)	28.9 ± 32.94
AST (IU/L)	30.2 ± 20.09
GGT (IU/L)	37.0 ± 63.69
FPG (mg/dL)	82.6 ± 8.34
ALB (g/dL)	4.26 ± 0.33
BUN (mg/dL)	14.5 ± 3.86
HDL-C (mg/dL)	44.7 ± 10.09
LDL-C (mg/dL)	116.5 ± 33.03
TG (mg/dL)	165.9 ± 106.01

Continuous variables were log transformed before analysis if not normally distributed. Mean ± SD. BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; WHR, waist-hip ratio; ALT, alanine aminotransferase, AST, asperate aminotransferase; GGT, gamma glutamyl transferase; FPG, fasting plasma glucose; ALB, albumin; BUN, blood urea nitrogen; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; TG, triglyceride.

Cell Culture and Electroporation

The human myelogenous leukemia cell line (K562) was purchased from American Type Culture Collection (ATCC) and the human heptacellular carcinoma cell line (Huh 7.5) was purchased from Korean Cell Line Bank (KCLB). K-562 cells were cultured in RPMI (GIBCO) and Huh 7.5 cells were cultured in DMEM-high glucose (GIBCO). All media were supplemented with 10% fetal bovine serum (FBS), penicillin (100 U/mL), and streptomycin (100 μg/mL). 1.5 × 10^6 K-562 cells and 4 × 10^5 Huh 7.5 cells were electroporated (Neon Transfection System, Invitrogen), with a 1:1:4 ratio of left TALEN:right TALEN:EGFP donor DNA (total, 6 μg).

Flow Cytometry

Suspension cells were collected while adherent cells were trypsinized and resuspended in PBS. Single-cell suspensions were analyzed and sorted using the FACSAria II (BD Biosciences).

Sorting Strong EGFP$^+$ Cells Containing TALEN-Induced Knockin

Untransfected cells were used as controls. Sorted cells were then used to obtain single-cell-derived clones by limiting dilution (0.25 cells per well of a 96-well plate). After 2 weeks, wells with cell populations from

Table 6. Disease Criteria for Finding No Disease Correlation Region

Disease Criteria	Trait	Type	Case	Control
T2D	glu0	CC	≤126	<110
Hypertension	HTN	CC	SBP ≥ 140 and DBP ≥ 90	90 ≤ SBP < 120, 60 ≤ DBP < 80
Osteoporosis	AS1_DT_cc	CC	T value ≤ −2.5	T value ≥ −1.0
	AS1_MT_cc	CC	T value ≤ −2.5	T value ≥ −1.0
Obesity	bmi	QT	–	–
Dyslipidemia	hdl	CC	<40	≤60
	tchl	CC	≤240	<200
	tg	CC	≤200 (except for more than 400)	<150
	ldl	CC	≤160	<100
Metabolic syndrome	MS_cc	CC	anyone with the presence of three or more following components: 1. central or abdominal obesity (measured by waist circumference) men: ≥102 cm; women: ≥88 cm (Asian: men: ≥90 cm; women: ≥80 cm) 2. fasting blood triglycerides ≥150 mg/dL 3. blood HDL cholesterol: men: <40 mg/dL; women: < 50 mg/dL 4. blood pressure ≥130/85 mmHg 5. fasting glucose ≥100 mg/dL	
Respiratory diseases	AS1_BrDs	CC	history of diagnosis for respiratory diseases	no
Arthritis	AS1_DgnArth	CC	history of diagnosis for degenerative arthritis	no
	AS1_RhmArth	CC	history of diagnosis for rheumatoid arthritis	no
Insomnia	AS1_Insm	CC	insomnia	no
Clinical test (blood and urea)	AS1_BCTRIA	CC	some bacteria was found	not found
	AS1_CRYSTAL1	CC	urine (16)-crystals were found	not found
	AS1_CRYSTAL2	CC	urine (16)-crystals: Ca.oxalate was found	not found
	AS1_CRYSTAL3	CC	urine (16)-crystals: triple phosphate was found	not found
	AS1_CRYSTAL4	CC	urine (16)-crystals: uric acid was found	not found
	AS1_CRYSTAL5	CC	urine (16)-crystals: Ca.phosphate was found	not found
	AS1_U_OTHR	QT	–	–
	AS1_VB12	QT	–	–
	AS1_FOLATE	QT	–	–
	AS1_VDRL	CC	venereal disease research laboratories test reactive (1:1)	not reactive
	AS1_FREET4	QT	–	–
	AS1_TSH	QT	–	–
	AS1_CD	QT	–	–
	AS1_PB	QT	–	–
	AS1_AL	QT	–	–
Body metrics	height	QT	–	–
	weight	QT	–	–
Lung function test	AS1_SP1_3	QT	–	–
	AS1_SP2_3	QT	–	–
	AS1_SP3_1	QT	–	–
Electrocardiogram	AS1_EKG	CC	EKG overall judgment: abnormal	normal
Chest X-ray	AS1_CH0	CC	chest X-ray overall opinion: abnormal	normal
Gender	sex	QT	–	–
Age	age	QT	–	–
Glasses	AS1_Glasses	CC	wearing glasses	no
Hearing aid	AS1_Acst	CC	wearing hearing aid	no

(*Continued on next page*)

Table 6. *Continued*

Disease Criteria	Trait	Type	Case	Control
Been in accidents	AS1_AccFq	QT	–	–
Tooth problem	AS1_Tooth	QT	–	–
Medical history	AS1_PdHn	CC	been diagnosed with head injury: yes	no
	AS1_PdUt	CC	been diagnosed with urinary tract infection: yes	no
	AS1_PdGt	CC	been diagnosed with gout: yes	no
	AS1_PdIm	CC	been diagnosed with erectile dysfunction: yes	no
Current medical diagnosis and treatment	AS1_TrtAr	CC	arthritis (degenerative, rheumatoid): yes	no
	AS1_TrtGt	CC	gout: yes	no

Complex diseases and complex disease-related traits were analyzed. Association study for cancer was not conducted because cancer patients were already excluded in the previous KARE GWAS. QT, quantitative trait; CC, case control; AS1_DT, distal radius T; AS1_MT, midshaft tibia T; AS1_DgnArth, degenerative arthritis; AS1_RhmArth, rheumatoid arthritis; AS1_BCTRIA, Urine_Bacteria; AS1_CRYSTAL1, Urine_Crystals; AS1_CRYSTAL2, Urine_Crystals_Ca.oxalate; AS1_CRYSTAL3, Urine_Crystals_Triple phosphate; AS1_CRYSTAL4, Urine_Crystals_Uric Acid; AS1_CRYSTAL5, Urine_Crystals_Ca.phosphate; AS1_U_OTHR, Urine_Others (found, not found); AS1_VB12, Vitamin B-12; AS1_FOLATE, folate; AS1_VDRL, venereal disease research laboratory test; AS1_FREET4, free T (thyroxine) 4; AS1_TSH, thyroid stimulation hormone; AS1_CD, cadmium; AS1_PB, plumbum; AS1_AL, aluminum; AS1_SP1, spirometry; AS1_CH0, chest X-ray (normal, abnormal); AS1_AccFq, frequency of accident; AS1_PdHn, diagnosis experience of external head injury; AS1_PdUt, positive diagnosis experience of urinary tract infection; AS1_PdGt, positive diagnosis experience of gout; AS1_PdIm, positive diagnosis experience of erectile dysfunction; AS1_TrtAr, treatment of arthritis; AS1_TrtGt, treatment of gout.

a single clone (round colony) were selected and expanded to perform flow cytometry using the BD FACSCanto system (MFI) and molecular analysis.

Clonal Analysis of Single Cells and Colonies

Before and after cell sorting, single cells were isolated using a mouth pipette under a microscope and transferred to PCR tubes. To obtain clonal populations of cells, sorted and unsorted cells were plated at a density of 1,000 cells per 100-mm plate, and colonies were manually picked after 2 weeks. For site-specific PCR analysis, the same donor-specific primer was used commonly among cells with locus-specific primer (AAVS1, CNVR7, and CNVR22). Then, PCR amplicons were cast on agarose gel and visualized by ethidium bromide staining.

Gene Expression Analysis

For gene expression analysis, total RNA was extracted from 1×10^6 cells using the TRIzol-chloroform method and reverse-transcribed with random primers according to the RT-&GO Mastermix (MP Biomedicals Asia Pacific, RTRAG100) manufacturer's protocol. We analyzed 200 ng cDNA from K562 or Huh 7.5 cells, respectively, in triplicate with TOPreal qRT-PCR 2X PreMIX (Enzynomics, RT500M) in a CFX96 real-time PCR detection system (Bio-Rad, C1000 real-time PCR thermal cycler). The relative expression level of each gene was B2M and YWHAZ expression (housekeeping gene controls) and represented as fold change relative to the mock-treated samples (calibrator).

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.omto.2019.07.001.

AUTHOR CONTRIBUTIONS

S.M., S.K., H.H.K., and B.-J.K. conceived and designed this study. Y.K.K., M.Y.H., and Y.J.K. conducted genome and epidemiological data analysis. E.-S.L. and K.D.A.-B. performed experiments. E.-S.L., S.M., H.H.K., and B.-J.K. wrote the manuscript. E-.S.L., S.M., S.K., N.S.H., H.H.K., and B.-J.K. revised the manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported by an intramural grant from the Korea National Institute of Health (2016-NI73001-00), the Institute of Basic Science (IBS-R026-D1), a National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (NRF-2018R1A5A2025079), and the Ministry of Education of the Republic of Korea (0668-20180123). The data used in this study were provided by the Korean Genome Analysis Project (4845-301), Korean Genome and Epidemiology Study (4851-302), and Korea Biobank Project (4851-307), which were supported by the Korea Centers for Disease Control and Prevention, Republic of Korea.

REFERENCES

1. Cideciyan, A.V., Hauswirth, W.W., Aleman, T.S., Kaushal, S., Schwartz, S.B., Boye, S.L., Windsor, E.A., Conlon, T.J., Sumaroka, A., Pang, J.J., et al. (2009). Human RPE65 gene therapy for Leber congenital amaurosis: persistence of early visual improvements and safety at 1 year. Hum. Gene Ther. *20*, 999–1004.

2. Aiuti, A., Cassani, B., Andolfi, G., Mirolo, M., Biasco, L., Recchia, A., Urbinati, F., Valacca, C., Scaramuzza, S., Aker, M., et al. (2007). Multilineage hematopoietic reconstitution without clonal selection in ADA-SCID patients treated with stem cell gene therapy. J. Clin. Invest. *117*, 2233–2240.

3. LeWitt, P., Schultz, L., Auinger, P., and Lu, M.; Parkinson Study Group DATATOP Investigators (2011). CSF xanthine, homovanillic acid, and their ratio as biomarkers of Parkinson's disease. Brain Res. *1408*, 88–97.

4. Wu, X., Li, Y., Crise, B., and Burgess, S.M. (2003). Transcription start regions in the human genome are favored targets for MLV integration. Science *300*, 1749–1751.

5. Mitchell, R.S., Beitzel, B.F., Schroder, A.R., Shinn, P., Chen, H., Berry, C.C., Ecker, J.R., and Bushman, F.D. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol. *2*, E234.

6. Schröder, A.R., Shinn, P., Chen, H., Berry, C., Ecker, J.R., and Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. Cell *110*, 521–529.

7. Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M.P., Wulffraat, N., Leboulch, P., Lim, A., Osborne, C.S., Pawliuk, R., Morillon, E., et al. (2003). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science *302*, 415–419.

8. Woods, N.B., Bottero, V., Schmidt, M., von Kalle, C., and Verma, I.M. (2006). Gene therapy: therapeutic gene causing lymphoma. Nature *440*, 1123.

9. Kim, H., and Kim, J.S. (2014). A guide to genome engineering with programmable nucleases. Nat. Rev. Genet. *15*, 321–334.

10. Sadelain, M., Papapetrou, E.P., and Bushman, F.D. (2011). Safe harbours for the integration of new DNA in the human genome. Nat. Rev. Cancer *12*, 51–58.

11. Smith, J.R., Maguire, S., Davis, L.A., Alexander, M., Yang, F., Chandran, S., ffrench-Constant, C., and Pedersen, R.A. (2008). Robust, persistent transgene expression in human embryonic stem cells is achieved with AAVS1-targeted integration. Stem Cells *26*, 496–504.

12. Ramachandra, C.J., Shahbazi, M., Kwang, T.W., Choudhury, Y., Bak, X.Y., Yang, J., and Wang, S. (2011). Efficient recombinase-mediated cassette exchange at the AAVS1 locus in human embryonic stem cells using baculoviral vectors. Nucleic Acids Res. *39*, e107.

13. Gierman, H.J., Indemans, M.H., Koster, J., Goetze, S., Seppen, J., Geerts, D., van Driel, R., and Versteeg, R. (2007). Domain-wide regulation of gene expression in the human genome. Genome Res. *17*, 1286–1295.

14. Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., et al. (2006). Copy number variation: new insights in genome diversity. Genome Res. *16*, 949–961.

15. Lombardo, A., Cesana, D., Genovese, P., Di Stefano, B., Provasi, E., Colombo, D.F., Neri, M., Magnani, Z., Cantore, A., Lo Riso, P., et al. (2011). Site-specific integration and tailoring of cassette design for sustainable gene transfer. Nat. Methods *8*, 861–869.

16. Kustikova, O., Fehse, B., Modlich, U., Yang, M., Düllmann, J., Kamino, K., von Neuhoff, N., Schlegelberger, B., Li, Z., and Baum, C. (2005). Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. Science *308*, 1171–1174.

17. Papapetrou, E.P., Lee, G., Malani, N., Setty, M., Riviere, I., Tirunagari, L.M., Kadota, K., Roth, S.L., Giardina, P., Viale, A., et al. (2011). Genomic safe harbors permit high β-globin transgene expression in thalassemia induced pluripotent stem cells. Nat. Biotechnol. *29*, 73–78.

18. Ogata, T., Kozuka, T., and Kanda, T. (2003). Identification of an insulator in AAVS1, a preferred region for integration of adeno-associated virus DNA. J. Virol. *77*, 9000–9007.

19. Bultmann, S., Morbitzer, R., Schmidt, C.S., Thanisch, K., Spada, F., Elsaesser, J., Lahaye, T., and Leonhardt, H. (2012). Targeted transcriptional activation of silent oct4 pluripotency gene by combining designer TALEs and inhibition of epigenetic modifiers. Nucleic Acids Res. *40*, 5368–5377.

20. Moon, S., Jung, K.S., Kim, Y.J., Hwang, M.Y., Han, K., Lee, J.Y., Park, K., and Kim, B.J. (2013). KGVDB: a population-based genomic map of CNVs tagged by SNPs in Koreans. Bioinformatics *29*, 1481–1483.

21. Cho, Y.S., Go, M.J., Kim, Y.J., Heo, J.Y., Oh, J.H., Ban, H.J., Yoon, D., Lee, M.H., Kim, D.J., Park, M., et al. (2009). A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. Nat. Genet. *41*, 527–534.

22. Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. (2002). Recent segmental duplications in the human genome. Science *297*, 1003–1007.

23. Schrider, D.R., and Hahn, M.W. (2010). Gene copy-number polymorphism in nature. Proc. Biol. Sci. *277*, 3213–3221.

24. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al.; Wellcome Trust Case Control Consortium (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704–712.

25. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. *17*, 1665–1674.