nature genetics

Exome sequencing identifies frequent mutation of *ARID1A* in molecular subtypes of gastric cancer

Kai Wang[1,7], Junsuo Kan[2,7], Siu Tsan Yuen[2], Stephanie T Shi[3], Kent Man Chu[4], Simon Law[4], Tsun Leung Chan[2], Zhengyan Kan[1], Annie S Y Chan[2], Wai Yin Tsui[2], Siu Po Lee[2], Siu Lun Ho[2], Anthony K W Chan[2], Grace H W Cheng[2], Peter C Roberts[5], Paul A Rejto[1], Neil W Gibson[1,6], David J Pocalyko[1], Mao Mao[1], Jiangchun Xu[1] & Suet Yi Leung[2]

Gastric cancer is a heterogeneous disease with multiple environmental etiologies and alternative pathways of carcinogenesis[1,2]. Beyond mutations in *TP53*, alterations in other genes or pathways account for only small subsets of the disease. We performed exome sequencing of 22 gastric cancer samples and identified previously unreported mutated genes and pathway alterations; in particular, we found genes involved in chromatin modification to be commonly mutated. A downstream validation study confirmed frequent inactivating mutations or protein deficiency of *ARID1A*, which encodes a member of the SWI-SNF chromatin remodeling family, in 83% of gastric cancers with microsatellite instability (MSI), 73% of those with Epstein-Barr virus (EBV) infection and 11% of those that were not infected with EBV and microsatellite stable (MSS). The mutation spectrum for *ARID1A* differs between molecular subtypes of gastric cancer, and mutation prevalence is negatively associated with mutations in *TP53*. Clinically, *ARID1A* alterations were associated with better prognosis in a stage-independent manner. These results reveal the genomic landscape, and highlight the importance of chromatin remodeling, in the molecular taxonomy of gastric cancer.

Recent studies using next-generation sequencing (NGS) have revealed an extensive repertoire of potential cancer-driving genes in several cancer types[3–17]. To further explore the genetic basis of gastric cancer, we performed whole-exome capture using the Agilent SureSelect Human All Exon kit followed by NGS on the Illumina Genome Analyzer IIx or HiSeq 2000 platforms to identify somatic mutations in 22 matched pairs of gastric cancer and normal tissue (**Supplementary Table 1**), with a mean depth of 116× and 91.4% of bases covered to at least 10× (**Supplementary Table 2**). Somatic mutations were predicted using algorithms described in the **Supplementary Note**, and an extensive subset was confirmed by Sequenom MassARRAY genotyping (**Supplementary Table 3**), giving a positive prediction rate of 96.8% (95% confidence interval (CI), 95.2–98.4%) for somatic single-nucleotide variation (SNV) and 97.9% (95% CI, 95.8–99.9%) for somatic insertion or deletion (indel) alterations.

Overall, 7,036 somatic mutations were detected in the 22 gastric cancer samples, of which 4,653 occurred in coding regions or essential splice sites (2,513 missense, 137 nonsense, 5 stop codon loss, 90 splice site, 855 indel and 1,053 synonymous mutations) (**Table 1** and **Supplementary Table 4**). Consistent with the known consequences of mismatch repair deficiency, the gastric cancer samples with MSI had an average of 31.61 somatic mutations (including both SNVs and indels) per megabase of DNA, whereas the MSS gastric cancer samples had an average of 3.29, a difference of approximately tenfold. Likewise, MSI gastric cancer samples had an average of 620 protein-altering somatic mutations, which was tenfold higher than the the number found in MSS gastric cancer samples (mean of 62). Although there was no significant difference in the nonsynonymous-to-synonymous (NS/S) ratio between the MSI and the MSS gastric cancers, we found a significantly higher NS/S ratio in EBV-infected gastric cancer samples (4.55 ± 0.97 compared to 2.76 ± 0.72 in non-EBV–infected gastric cancer samples, $P < 0.012$) and in poorly differentiated gastric cancers (3.52 ± 0.94 compared to 2.56 ± 0.78 in well- and moderately differentiated gastric cancers, $P < 0.014$), suggestive of a higher positive selection pressure for driver mutations in these subgroups[18,19]. Of note, the NS/S ratio for the EBV-infected gastric cancer samples (range 3.64–5.56) is among the highest for solid organ cancers (range 0.97–3.5) reported to date[3,8–10,12,17]. C-to-T transitions were the most common mutation (51%) across all gastric cancers, with 68% involving CG dinucleotides. The MSI gastric cancers also had a distinctly high rate of T-to-C transitions (30%) (**Supplementary Table 5**). The observed mutation incidence for MSS gastric cancers (3.29 per megabase) is higher than our previous results from sequencing the kinome of most other cancers (0.74–1.85 per megabase for renal, colorectal and ovarian cancers) but lower than in lung cancers (4.21 per megabase)[20]. The observed mutation spectrum

[1]Oncology Research Unit, Pfizer Worldwide Research and Development, La Jolla, California, USA. [2]Department of Pathology, The University of Hong Kong, Queen Mary Hospital, Pokfulam, Hong Kong. [3]External Research Solutions, Pfizer Worldwide Research and Development, La Jolla, California, USA. [4]Department of Surgery, The University of Hong Kong, Queen Mary Hospital, Pokfulam, Hong Kong. [5]Research Embedded Business Technology, Pfizer Worldwide Research and Development, La Jolla, California, USA. [6]Present address: Regulus Therapeutics, San Diego, California, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to S.Y.L. (suetyi@hku.hk) or J.X. (jiangchun.xu@pfizer.com).

Received 18 May; accepted 23 September; published online 30 October 2011; doi:10.1038/ng.982

Table 1 Summary of somatic mutation types and prevalence in 22 gastric cancers

	Coding regions and essential splice sites									Other non-coding regions	
	SNV										
Sample	Synonymous	Missense	Stop gained	Stop lost	Essential splice site	Indel	Total	Mutation per Mb DNA	NS/S	SNV	Indel
MSS samples											
pfg001T	12	44	5	0	1	2	64	2.77	4.08	21	3
pfg002T	6	18	1	0	0	1	26	1.08	3.17	7	0
pfg003T	6	17	2	0	0	1	26	1.11	3.17	4	2
pfg005T	0	4	0	0	0	0	4	0.17	–	0	1
pfg006T	24	41	4	0	2	3	74	3.04	1.88	17	3
pfg007T	23	38	3	0	0	4	68	2.80	1.78	20	3
pfg009T[a]	16	85	4	0	2	7	114	4.65	5.56	19	3
pfg010T	18	34	0	0	2	5	59	2.41	1.89	16	3
pfg011T	20	65	7	0	3	8	103	3.99	3.60	20	6
pfg014T	45	90	7	0	3	5	150	5.86	2.16	40	8
pfg015T	25	57	4	0	3	4	93	3.64	2.44	18	2
pfg018T[a]	11	44	5	0	0	4	64	2.50	4.45	16	3
pfg020T	17	56	4	0	2	5	84	3.27	3.53	18	2
pfg021T	9	25	3	0	2	0	39	1.56	3.11	4	1
pfg022T[a]	11	38	2	0	0	1	52	2.00	3.64	17	2
pfg024T	28	68	1	0	1	4	102	3.92	2.46	30	2
pfg025T	18	65	5	0	1	3	92	3.58	3.89	14	5
pfg029T	67	159	8	0	15	14	263	10.21	2.49	57	5
Total MSS	356	948	65	0	37	71	1,477	3.29	2.85	338	54
MSI samples											
pfg008T	303	643	20	4	33	226	1,229	51.62	2.20	278	439
pfg016T	217	538	27	1	12	250	1,045	40.92	2.61	182	450
pfg017T	29	82	8	0	4	122	245	9.61	3.10	28	212
pfg019T	148	302	17	0	4	186	657	25.62	2.16	95	307
Total MSI	697	1,565	72	5	53	784	3,176	31.61	2.36	583	1,408
Overall total	1,053	2,513	137	5	90	855	4,653	8.48	2.52	921	1,462

Mutation per Mb DNA was calculated by dividing the total number of somatic SNVs and indels overlapping with the coding regions and essential splice sites by the total number of coding bases sufficiently covered (≥3× in tumor and ≥10× in matched normal samples) by sequencing data.
[a]EBV-infected gastric cancer samples.

is similar to our previous report for both MSS and MSI gastric cancers[20]. In the current study, we have generated a list of 2,890 genes harboring protein-altering somatic mutations in gastric cancer, 59 of which were reported in the kinome study[20] and in the Catalogue of Somatic Mutations in Cancer (COSMIC)[21]. Thus, 2,831 new genes have been discovered to be mutated in gastric cancer (**Supplementary Table 6**).

Eighty-two genes were mutated in three or more samples and 447 genes in two or more samples (**Supplementary Table 7**). Using a driver-gene score[22,23] (described in the **Supplementary Note**), 20 genes were identified as candidate drivers in our gastric cancer cohort at a false discovery rate (FDR) of ≤0.2 (**Table 2**). These genes included known drivers of gastric carcinogenesis, such as *TP53* (ref. 24; mutated in 36% of samples), *PTEN*[25] (27%) and *CTNNB1* (ref. 26; 9%), as well as genes reported in COSMIC to be mutated in gastric cancer, such as *TTK*[27] (mutated in 18% of samples) and *ACVR2A*[28] (18%). In addition, we identified many genes previously known to associate with gastric cancer with slightly higher FDRs, including *PIK3CA*[29] (mutated in 14% of samples), *APC*[30] (14%) and *CDH1* (refs. 31,32; 9%) (**Supplementary Table 7**). Our candidate drivers also included highly penetrant variations in genes not previously reported to be mutated in gastric cancer, including *ARID1A* (6 of 22 gastric cancer samples or 27%) and other genes with diverse cellular functions of potential importance in cancer, such as *FMN2* (involved in cytoskeletal organization and cell polarity) and *SEMA3E* (involved in axon guidance)[33].

We searched for over-represented molecular pathways among genes with protein-altering somatic mutations of biological significance (**Supplementary Tables 8**, **9** and **Supplementary Note**). Chromatin modification and cell junction organization were the pathways with the most significant enrichment of mutated genes. Frequent mutations of chromatin remodeling genes, such as *ARID1A*, *PBRM1*, *MEN1* and *DAXX*, were recently discovered in other cancers, but their involvement in gastric cancer had not been reported[4,5,7,11]. We found that, in addition to *ARID1A*, other members of the SWI-SNF complex (*ARID1B, PBRM1* and *SMARCC1*), ISWI complex (*SMARCA1*) and NuRD complex (*CHD3*, *CHD4* and *MBD2*), as well as other genes encoding histone-modifying proteins (*SIRT1* and *SETD2*), were also mutated, in total affecting 59% of gastric cancers. Moreover, a single gastric cancer sample could have mutations in multiple chromatin remodeling genes involving different complexes (**Supplementary Table 10**). Overall, 59% of gastric cancer samples had mutations in at least one gene affecting cell junction organization, including *CDH1* and other cadherin family members, consistent with the infiltrative nature of gastric cancer and its great tendency toward the loss of cell-to-cell adhesion. Genes involved in cell cycle regulation were mutated in 77% of gastric cancers, including *TP53*, *PTEN* and *TTK*. Other core signaling pathways frequently mutated in gastric cancer included the Wnt-BMP-TGFβ, axon guidance, MAPK, DNA replication, focal adhesion, ERBB, ATR-BRCA and Rb pathways, many of which may have therapeutic implications.

Table 2 Candidate driver genes in gastric cancer predicted from exome sequencing of 22 gastric cancers

Gene symbol	Name	Size	N_S	N_T	N_{SNV}^{MSS}	N_{SNV}^{MSI}	N_{indel}^{MSS}	N_{indel}^{MSI}	*P*	*Q* value	*Q* score
TP53	Cellular tumor antigen p53	1,182	8	8	8	0	0	0	0.0000	0.0000	8.26
PTEN	PIP_3 phosphatase and dual-specificity protein phosphatase (PTEN)	1,212	6	8	2	2	0	4	0.0000	0.0000	4.43
ARID1A	AT-rich interactive domain–containing protein 1A	6,414	6	7	2	0	1	4	0.0008	0.0937	1.03
RPL22	60S ribosomal protein L22	387	3	3	0	0	0	3	0.0013	0.1202	0.92
TTK	Dual specificity protein kinase TTK	2,511	4	5	0	0	0	5	0.0019	0.1206	0.92
FMN2	Formin-2	5,396	4	5	4	1	0	0	0.0018	0.1206	0.92
SPRR2B	Small proline-rich protein 2B	219	2	2	2	0	0	0	0.0024	0.1206	0.92
PTN	Pleiotrophin precursor	507	2	2	1	0	1	0	0.0024	0.1206	0.92
ACVR2A	Activin receptor type-2A precursor	1,542	4	4	0	0	0	4	0.0032	0.1339	0.87
PMS2L3	Postmeiotic segregation increased 2-like protein 3	515	2	2	0	1	1	0	0.0033	0.1339	0.87
DNAH7	Dynein heavy chain 7, axonemal	12,064	7	8	5	2	0	1	0.0044	0.1446	0.84
TTN	Titin isoform novex-3	82,794	5	7	3	4	0	0	0.0039	0.1446	0.84
FSCB	Fibrous sheath CABYR-binding protein	2,478	3	3	3	0	0	0	0.0048	0.1446	0.84
CTNNB1	Catenin beta-1	2,346	2	3	3	0	0	0	0.0049	0.1446	0.84
SEMA3E	Semaphorin-3E precursor	2,328	3	3	0	2	1	0	0.0057	0.1587	0.80
MCHR1	Melanin-concentrating hormone receptor 1	1,269	3	3	0	3	0	0	0.0062	0.1627	0.79
SPANXN2	Sperm protein associated with the nucleus on the X chromosome N2	543	2	2	2	0	0	0	0.0066	0.1641	0.79
METTL3	N6-adenosine-methyltransferase 70 kDa subunit	1,518	2	3	0	3	0	0	0.0078	0.1845	0.73
EIF3A	Eukaryotic translation initiation factor 3 subunit A	4,047	3	4	0	4	0	0	0.0089	0.1947	0.71
EPB41L3	Band 4.1-like protein 3	3,165	2	3	3	0	0	0	0.0091	0.1947	0.71

Numbers of somatic mutations shown include only protein-altering mutations. Size, interrogated size of a gene, obtained by counting the coding bases targeted by the SureSelect capture baits; N_S, number of cancer samples in which a gene is mutated; N_T, total number of mutations; N_{SNV}^{MSS}, number of SNV mutations in MSS cancers; N_{SNV}^{MSI}, number of SNV mutations in MSI cancers; N_{indel}^{MSS}, number of indel mutations in MSS cancers; N_{indel}^{MSI}, number of indel mutations in MSI cancers. The derivation of driver gene statistics (*P* values, *Q* values and *Q* scores) is described in the **Supplementary Note**.

The frequent mutation of genes encoding chromatin remodelers in gastric cancer and, specifically, the discovery of *ARID1A* as one of the top candidate driver genes in gastric cancer prompted us to study *ARID1A* in greater detail. *ARID1A* was recently discovered to be a driver for ovarian clear cell carcinoma[7,11]. Because of the incomplete coverage of the *ARID1A* gene in exome sequencing, we resequenced all coding exons of the discovery cohort as well as an additional 87 gastric cancer samples by Sanger sequencing. The two cohorts together encompass 109 gastric cancers, including 23 MSI[34,35], 15 MSS, EBV-infected[36,37] and 71 MSS, non-EBV–infected samples (**Supplementary Tables 1** and **11**). We found a striking difference in the somatic mutation rate of *ARID1A* across the different molecular subtypes of gastric cancer, with very high incidence in the MSI types (78%, 18/23, $P < 0.001$ compared to MSS, non-EBV–infected) and the MSS, EBV-infected gastric cancers (47%, 7/15, $P = 0.002$ compared to MSS, non-EBV–infected) and lower incidence in MSS, non-EBV–infected gastric cancers (10%, 7/71) (**Fig. 1** and **Supplementary Table 12**). Among the 32 gastric cancer samples containing *ARID1A* mutations, 14 samples had two mutations and 18 had a single mutation, giving a total of 46 *ARID1A* mutations, out of which 39 (85%) were truncating mutations (**Fig. 2**, **Supplementary Table 13** and **Supplementary Fig. 1**). The majority (75%, 24/32) of gastric cancers with *ARID1A* mutations had either a loss of or substantially lower protein expression compared to gastric cancers without *ARID1A* mutation ($P < 0.001$), as determined by immunostaining (**Fig. 1**, **Supplementary Tables 12**, **13** and **Supplementary Fig. 2**). Of note, six gastric cancer samples showed absent or weak protein expression despite the lack of detectable *ARID1A* mutations, suggesting that other mechanisms may contribute to ARID1A inactivation. Combining mutation and protein expression data, ARID1A alterations (either through mutation or reduced protein levels) were detected in 73% (11/15) of EBV-infected gastric cancer samples, at levels comparable to the alteration rate in MSI gastric cancers (83%, 19/23) and substantially higher than in the MSS, non-EBV–infected subtype (11%, 8/71) (**Fig. 1** and **Supplementary Table 12**). Gastric cancers with ARID1A alterations were less likely to harbor mutations in the *TP53* gene. *TP53* mutations were observed in 21% (8/38) of gastric cancer samples with ARID1A alterations but in 52% (37/71) of samples without alterations ($P = 0.002$; **Fig. 1** and **Supplementary Table 12**). Of note, *TP53* mutation was uncommon in both MSI (17.4%, 4/23) and EBV-infected gastric cancers (6.7%, 1/15). Gastric cancer samples with ARID1A alterations showed a trend of prolonged, recurrence-free survival ($P = 0.058$, log-rank test) (**Supplementary Fig. 3**). More importantly, in multivariate analysis incorporating tumor stage, Lauren's tumor type, MSI status and ARID1A alteration status, only advanced tumor stage

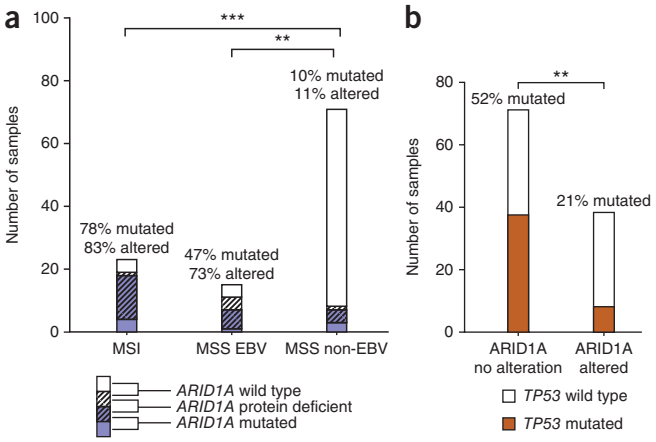

Figure 1 Relationship of ARID1A alterations (mutation or protein deficiency) with molecular subtypes of gastric cancer. (**a**) Incidence of *ARID1A* mutation and protein deficiency in different molecular subtypes of gastric cancer. Blue color indicates samples with *ARID1A* mutation, and shaded area indicates samples with protein deficiency, as determined by immunohistochemistry. $^{**}P < 0.01$, $^{***}P < 0.001$ for both ARID1A mutation and alteration (see **Supplementary Table 12** for detail). (**b**) *TP53* mutation rate in gastric cancer samples with or without ARID1A alterations.

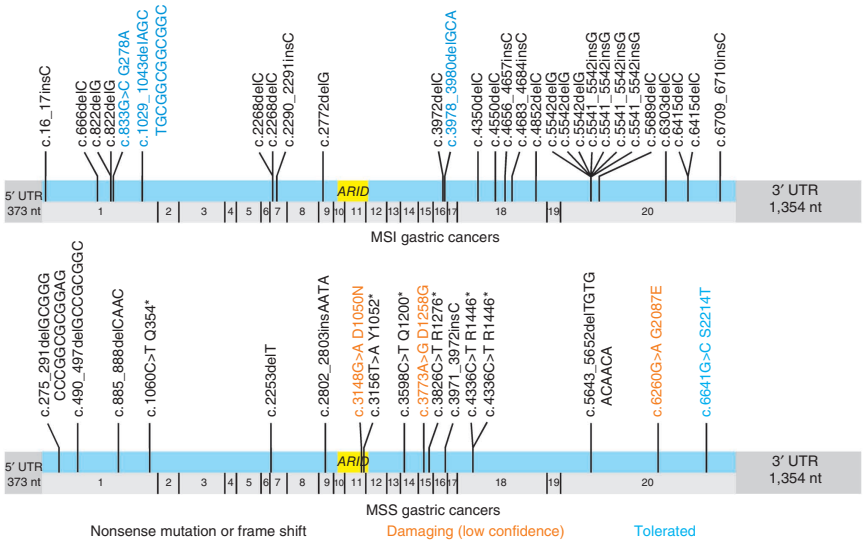

Figure 2 Difference in the mutation spectrum of *ARID1A* between molecular subtypes of gastric cancer. Individual exons of the *ARID1A* gene are represented as numbered gray boxes. Mutations detected by sequencing the coding region of *ARID1A* in 23 MSI and 86 MSS gastric cancers are shown. cDNA and peptide positions are based on the ENST00000324856 transcript. Colors of text correspond to functional impact prediction by SIFT, as indicated at the bottom.

(adjusted hazard ratio stage IV compared to stage I, 13.8; stage III compared to stage I, 9.25; $P = 0.001$) and absence of ARID1A alterations (adjusted hazard ratio 3.09, $P = 0.029$) were independent variables that could predict early recurrence (**Supplementary Table 14**).

The mutation spectrum of *ARID1A* was strikingly different between the MSI and MSS subtypes (**Fig. 2**). In the MSI gastric cancer samples, 97% (28/29) of the mutations were indels, mostly involving short mononucleotide repeats of C or G (89%, 25/28). Specifically, one single G7 tract located in exon 20 was mutated in 26% of MSI gastric cancers. For the MSS gastric cancer samples (both EBV infected and non–EBV infected), 59% (10/17) of the mutations were SNVs with 6 nonsense and 4 missense mutations. Only 7 mutations were indels, with 1 involving a mononucleotide repeat sequence. *ARID1A* contains many short repeats of 4–7 mononucleotides in its coding region. The high rate of somatic indels at these repeats in MSI gastric cancer samples is not likely to be caused simply by a high background mutation rate. Compared with the global background mutation rate of somatic indels at mononucleotide tracts of similar length in MSI gastric cancers, the mutation rate in *ARID1A* is 12- to 61-fold higher ($P < 0.0001$; **Supplementary Fig. 4** and **Supplementary Table 15**), consistent with a driver gene being targeted by the MSI mechanism[38]. The overall mutation rate of *ARID1A* in MSI gastric cancer (78%) is comparable to that of well-established and functionally validated driver genes inactivated by MSI, such as *TGFBR2* (ref. 39).

This study revealed for the first time the diverse gene and pathway alterations in gastric cancer and the importance of chromatin remodeling genes, with frequent mutations in *ARID1A* and other associated factors affected 59% of gastric cancer samples in total. Consistent with our findings, loss of protein expression of ARID1A has been noted in 11–14% of gastric cancer samples by immunostaining[40,41], and homozygous deletion of *ARID1A* has been detected in a gastric cancer cell line[4]. Interestingly, we noted an inverse relationship between the two most frequently mutated gastric cancer genes *ARID1A* and *TP53*, suggesting that they may drive alternative subsets of gastric cancer. Mutation of *TP53* is rare in other cancers occurring in the ovary[7,11], endometrium[41], kidney[4] or endocrine pancreas[5], in which mutation of chromatin-modifying genes is common. Likewise, *TP53* mutation is uncommon in both MSI and EBV-infected gastric cancers, two subtypes with frequent *ARID1A* mutation. Thus, mutation of genes encoding chromatin-remodeling enzymes may constitute an alternative pathway of carcinogenesis independent of *TP53* that drives cancer development through epigenetic modification.

Comparison of the *ARID1A* mutation spectrum between gastric and ovarian cancers[7] revealed ten overlapping mutations, all of which are indels involving G7 or C6 repeats. These mutations are potentially caused by MSI, as a meta-analysis showed that 11% of clear cell ovarian cancers are mismatch repair deficient[42]; thus, mutation of *ARID1A* may be prevalent in MSI tumors arising in other organ sites. The high rate of ARID1A alteration (73%) in gastric cancers with EBV infection is intriguing. It would be interesting to examine other EBV-associated cancer types to see if the association with *ARID1A* mutation is gastric specific. As in ovarian clear cell carcinoma[7], we noted frequent two-hit inactivation of *ARID1A*, supporting its role as a tumor suppressor gene. For gastric cancer samples with a single mutation, protein loss was commonly observed, indicating that other mechanisms may contribute to inactivation. Moreover, haploinsufficiency of *ARID1A* may already contribute to carcinogenesis, as mice heterozygous for *ARID1A* mutation are embryonic lethal[43]. ARID1A functions as a component of the SWI-SNF complex and can repress gene expression, including that of *MYC* and of genes regulated by the E2F transcription factors[44–46]. More interestingly, a previous shRNA screen revealed that knockdown of *ARID1A* in Jurkat cells led to inhibition of Fas-mediated apoptosis[47]. One histological feature unique to MSI[35,48] and EBV-infected gastric cancers[2,36,37] is abundant tumor-infiltrating lymphocytes. Thus, it is tempting to speculate that selection for mutated *ARID1A* in these gastric cancer subtypes may be due to the ability of this mutation to confer resistance to Fas-mediated apoptosis and thereby promote immune evasion. Our preliminary analysis shows that the observed better prognosis in individuals with gastric cancer with ARID1A alterations may not be attributable solely to a link between ARID1A alteration and MSI status. ARID1A alteration is predictive of disease-free survival even after controlling for other clinicopathological variables, including tumor stage and MSI status, suggesting that ARID1A alterations may define a molecular subgroup of gastric cancer with a unique mechanism of carcinogenesis leading to distinct clinical behavior. Finally, our discovery of frequent ARID1A alterations in two specific molecular subtypes of gastric cancer underscores the importance of a comprehensive understanding of pathway-specific molecular changes to support the development of targeted molecular therapy and raises the possibility that specific epigenetic therapy targeting alterations in ARID1A or other chromatin-modifying enzymes may be useful in treating gastric cancer.

URLs. European Genome-Phenome Archive, www.ebi.ac.uk/ega/; COSMIC, http://www.sanger.ac.uk/genetics/CGP/cosmic/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession numbers. Exome sequencing data has been deposited in the European Genome-Phenome Archive with the study ID EGAS00001000153.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank L. Ng and M.K. Ng for technical assistance and clinicians in the Hong Kong Hospital Authority for clinical care. We thank Illumina for performing whole-exome sequencing.

AUTHOR CONTRIBUTIONS

D.J.P., P.A.R., N.W.G., M.M., J.X., S.T.Y. and S.Y.L. conceived of the study. S.T.Y., S.Y.L., J.X. and M.M. directed the study. K.W., S.T.S., P.A.R., J.X., M.M., S.T.Y. and S.Y.L. contributed to the project design. K.W., Z.K. and G.H.W.C. performed the bioinformatics data analysis. J.K., T.L.C., A.S.Y.C., W.Y.T., S.P.L., S.L.H. and A.K.W.C. performed experiments on *ARID1A* mutation and other molecular analysis. K.M.C. and S.L. contributed samples, data and comments on the manuscript. P.C.R. contributed to data management. K.W., J.K., S.T.Y., M.M., J.X. and S.Y.L. analyzed and interpreted data and wrote the manuscript with the assistance and final approval from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Lauwers, G.Y. *et al.* Gastric carcinoma. in *WHO Classification of Tumours of the Digestive System* (eds. Bosman, F.T., Carneiro, F., Hruban, R.H. & Theise, N.D.) 48–58 (IARC, Lyon, 2010).
2. Osato, T. & Imai, S. Epstein-Barr virus and gastric carcinoma. *Semin. Cancer Biol.* **7**, 175–182 (1996).
3. Wei, X. *et al.* Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat. Genet.* **43**, 442–446 (2011).
4. Varela, I. *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**, 539–542 (2011).
5. Jiao, Y. *et al.* DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. *Science* **331**, 1199–1203 (2011).
6. Chapman, M.A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–472 (2011).
7. Wiegand, K.C. *et al.* ARID1A mutations in endometriosis-associated ovarian carcinomas. *N. Engl. J. Med.* **363**, 1532–1543 (2010).
8. Pleasance, E.D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
9. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
10. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477 (2010).
11. Jones, S. *et al.* Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science* **330**, 228–231 (2010).
12. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
13. Shah, S.P. *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
14. Mardis, E.R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361**, 1058–1066 (2009).
15. Ley, T.J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
16. Campbell, P.J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40**, 722–729 (2008).
17. Totoki, Y. *et al.* High-resolution characterization of a hepatocellular carcinoma genome. *Nat. Genet.* **43**, 464–469 (2011).
18. Hughes, A.L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170 (1988).
19. Li, W.H. & Gojobori, T. Rapid evolution of goat and sheep globin genes following gene duplication. *Mol. Biol. Evol.* **1**, 94–108 (1983).
20. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
21. Forbes, S.A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
22. Kan, Z. *et al.* Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869–873 (2010).
23. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
24. Uchino, S. *et al.* p53 mutation in gastric cancer: a genetic model for carcinogenesis is common to gastric and colorectal cancer. *Int. J. Cancer* **54**, 759–764 (1993).
25. Wang, J.Y. *et al.* Mutation analysis of the putative tumor suppressor gene PTEN/MMAC1 in advanced gastric carcinomas. *Virchows Arch.* **442**, 437–443 (2003).
26. Park, W.S. *et al.* Frequent somatic mutations of the beta-catenin gene in intestinal-type gastric cancer. *Cancer Res.* **59**, 4257–4260 (1999).
27. Ahn, C.H., Kim, Y.R., Kim, S.S., Yoo, N.J. & Lee, S.H. Mutational analysis of TTK gene in gastric and colorectal cancers with microsatellite instability. *Cancer Res. Treat.* **41**, 224–228 (2009).
28. Hempen, P.M. *et al.* Evidence of selection for clones having genetic inactivation of the activin A type II receptor (ACVR2) gene in gastrointestinal cancers. *Cancer Res.* **63**, 994–999 (2003).
29. Li, V.S. *et al.* Mutations of PIK3CA in gastric adenocarcinoma. *BMC Cancer* **5**, 29 (2005).
30. Nakatsuru, S. *et al.* Somatic mutation of the APC gene in gastric cancer: frequent mutations in very well differentiated adenocarcinoma and signet-ring cell carcinoma. *Hum. Mol. Genet.* **1**, 559–563 (1992).
31. Becker, K.F. *et al.* E-cadherin gene mutations provide clues to diffuse type gastric carcinomas. *Cancer Res.* **54**, 3845–3852 (1994).
32. Guilford, P. *et al.* E-cadherin germline mutations in familial gastric cancer. *Nature* **392**, 402–405 (1998).
33. Chedotal, A., Kerjan, G. & Moreau-Fauvarque, C. The brain within the tumor: new roles for axon guidance molecules in cancers. *Cell Death Differ.* **12**, 1044–1056 (2005).
34. Leung, S.Y. *et al.* hMLH1 promoter methylation and lack of hMLH1 expression in sporadic gastric carcinomas with high-frequency microsatellite instability. *Cancer Res.* **59**, 159–164 (1999).
35. Leung, S.Y. *et al.* Microsatellite instability, Epstein-Barr virus, mutation of type II transforming growth factor beta receptor and BAX in gastric carcinomas in Hong Kong Chinese. *Br. J. Cancer* **79**, 582–588 (1999).
36. Yuen, S.T. *et al.* In situ detection of Epstein-Barr virus in gastric and colorectal adenocarcinomas. *Am. J. Surg. Pathol.* **18**, 1158–1163 (1994).
37. Chen, X. *et al.* Variation in gene expression patterns in human gastric cancers. *Mol. Biol. Cell* **14**, 3208–3215 (2003).
38. Woerner, S.M., Kloor, M., von Knebel Doeberitz, M. & Gebert, J.F. Microsatellite instability in the development of DNA mismatch repair deficient tumors. *Cancer Biomark.* **2**, 69–86 (2006).
39. Markowitz, S. *et al.* Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science* **268**, 1336–1338 (1995).
40. Wiegand, K.C. *et al.* Loss of BAF250a (ARID1A) is frequent in high-grade endometrial carcinomas. *J. Pathol.* **224**, 328–333 (2011).
41. Guan, B. *et al.* Mutation and Loss of Expression of ARID1A in Uterine Low-grade Endometrioid Carcinoma. *Am. J. Surg. Pathol.* **35**, 625–632 (2011).
42. Murphy, M.A. & Wentzensen, N. Frequency of mismatch repair deficiency in ovarian cancer: a systematic review. *Int. J. Cancer* **129**, 1914–1922 (2011).
43. Gao, X. *et al.* ES cell pluripotency and germ-layer formation require the SWI/SNF chromatin remodeling component BAF250a. *Proc. Natl. Acad. Sci. USA* **105**, 6656–6661 (2008).
44. Nagl, N.G. Jr., Zweitzig, D.R., Thimmapaya, B., Beck, G.R. Jr. & Moran, E. The c-myc gene is a direct target of mammalian SWI/SNF-related complexes during differentiation-associated cell cycle arrest. *Cancer Res.* **66**, 1289–1293 (2006).
45. Nagl, N.G. Jr., Wang, X., Patsialou, A., Van Scoy, M. & Moran, E. Distinct mammalian SWI/SNF chromatin remodeling complexes with opposing roles in cell-cycle control. *EMBO J.* **26**, 752–763 (2007).
46. Inoue, H. *et al.* Target genes of the largest human SWI/SNF complex subunit control cell growth. *Biochem. J.* **434**, 83–92 (2011).
47. Luo, B. *et al.* Highly parallel identification of essential genes in cancer cells. *Proc. Natl. Acad. Sci. USA* **105**, 20380–20385 (2008).
48. dos Santos, N.R., Seruca, R., Constancia, M., Seixas, M. & Sobrinho-Simoes, M. Microsatellite instability at multiple loci in gastric carcinoma: clinicopathologic implications and prognosis. *Gastroenterology* **110**, 38–44 (1996).

ONLINE METHODS

Sample preparation, MSI and EBV testing. The use of archival frozen human gastric samples for this study was approved by the Institutional Review Board of the University of Hong Kong and the Hospital Authority Hong Kong West Cluster. Because samples were archival in nature, the IRBs waived the need for informed consent, in compliance with international guidelines. Frozen tumor and non-neoplastic gastric tissues were collected from gastrectomy specimens from Queen Mary Hospital, The University of Hong Kong. DNA was extracted from frozen tumor tissue after macrodissection, and the tumor content was confirmed to be above 70% by cryostat sectioning for samples submitted for exome sequencing and above 50% for the validation cohort. DNA was also extracted from paired unaffected gastric tissue distant from the tumor site that was confirmed to be tumor free by prior histological examination of cryostat sections. Microsatellite instability analysis was performed using the standard method as previously described[35,49], with the addition of several loci, including *SLC7A8*, *TMPPB5* and *ZNF2*. At least five loci were analyzed in each case, including both dinucleotide and mononucleotide repeats. Those with high-level microsatellite instability (at least 40% of markers positive) were considered MSI. Those with low-level or no microsatellite instability (<40% of markers positive) were considered MSS. The presence of EBV in cancer cells was detected by *in situ* hybridization for EBV-encoded RNA (EBER) as described previously[36,37] or using the Ventana INFORM EBER Probe (Ventana Medical Systems) according to the manufacturer's protocol. The RNA Positive Control Probe (Ventana Medical Systems) was used to ensure RNA integrity. *Helicobacter pylori* infection status was defined by its presence in the gastrectomy specimen or endoscopic gastric biopsy before surgery.

Immunohistochemistry for ARID1A. Immunostaining was performed using a rabbit polyclonal antibody recognizing human ARID1A (Sigma-Aldrich HPA005456, dilution 1 in 200). After heat-mediated antigen retrieval and primary antibody incubation, signal was detected using the LabVision UltraVision kit (Thermo Scientific) and developed using diaminobenzidine counterstained with Mayer's hematoxylin. Staining was performed on a tissue microarray containing the gastric cancer tissues, and histological images were obtained with the ScanScope CS digital scanning system (Aperio Technologies). Nuclear staining of cancer cells was graded in comparison to normal cells (including infiltrating lymphocytes, fibroblasts and normal gastric glandular cells). ARID1A is strongly and uniformly expressed in normal cells, including gastric epithelial cells, lymphocytes and fibroblasts. ARID1A expression in gastric cancer cells was graded as 'normal' if staining intensity was similar to that in normal cells, 'weak' if staining intensity was substantially weaker than in normal cells or 'loss' if nuclear staining was absent, in which case the normal cells served as an internal positive control. Grading was performed independently by two pathologists (S.T.Y. and S.L.H.) who were blinded to the mutation results. Concordance rates between the two pathologists were high, and the few discrepancies were reviewed and consensus grading was assigned.

Exome capture, library construction and sequencing. Exome capture was performed by Illumina using Agilent SureSelect in-solution target enrichment technology (Agilent Technologies). Libraries were constructed following the Illumina Paired-End Sequencing Library Preparation Protocol version 1.0.1 from the SureSelect Human All Exon kit, with an added gel purification step for insert size selection. The kit contains a pool of RNA-based 120-mer capture oligomers (or baits) targeting 37,640,396 bases of 165,637 consensus coding sequence exons and their flanking regions. Sequencing was performed in two phases. In the first phase, libraries were constructed for ten pairs of tumor and matched normal samples with an insert size of ~300 bp and sequenced using two lanes on an Illumina Genome Analyzer II× sequencer. In the second phase, 12 additional pairs of samples were sequenced on one lane of the newer HiSeq 2000 sequencer and with a targeted insert size of ~180 bp. All sequencing was run with paired-end 75-bp reads and was performed according to Illumina's standard protocol. Library construction and sequencing of DNA from subject pfg011 were done in duplicate to assess the reproducibility of the technology and to serve as controls for the somatic mutation calling algorithm (see **Supplementary Note**). Reads from both replicates were combined in the final analysis. On average, ~158.5 million purity-filtered reads were generated for each sample. The mean percentage of duplicate reads due to PCR and optical artifacts was 6.3% in our data set. After removing these duplicate reads, ~133.5 million uniquely mapped reads were obtained for each sample. On average, 59.3% of reads in each sample had at least 50% overlap with any targeted region ± 100 bp in the SureSelect whole-exome bait library. The targeted regions in each sample were sequenced to an average depth of 115.8×, with ~97.8% of the targeted regions covered ≥1×, ~91.4% ≥10×, ~77.6% ≥30× and ~65.7% ≥50×. We have also genotyped ~1.1 million markers using Illumina HumanOmni1-Quad BeadChips on the same set of samples. The average concordance between the array-based and sequencing-based genotyping calls was 99.5%. A detailed summary of sequencing statistics for all samples can be found in **Supplementary Table 2**.

Mutation confirmation using the Sequenom MassARRAY system. Sequenom MassARRAY assays were performed by Sequenom. Assays were designed by using ProxSNP and PreXTEND analysis and validated by finding successful PCR and primer extension in Coriell DNA. Sample analysis was carried out using Sequenom iPLEX PRO chemistry. All Typer 4.0.20 calls were manually confirmed by examining the spectra for each assay and sample. Peaks for two alleles were checked against the background of each well, and a mutation call was confirmed if the peak was unique to that allele.

Mutation analysis for *ARID1A* and *TP53* in gastric cancers and matched normal tissues by Sanger sequencing. The coding exons of *ARID1A* were assessed by Sanger sequencing using primers similar to those previously published[11] with minor modifications (**Supplementary Table 16**). The discovery cohort of 22 gastric cancer samples was also resequenced because of incomplete coverage in some regions of *ARID1A* in exome sequencing. An additional 87 gastric cancers were also sequenced. All mutations identified in tumors were confirmed by independent PCR and Sanger sequencing in the specific tumors and their paired normal tissue to determine their somatic nature. Sequencing was performed using the ABI BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems). The sequence chromatograms were visually inspected with assistance from Mutation Surveyor.

For analysis of the *TP53* gene, exons 5 to 9 that include the mutation hotspots were sequenced using the primers listed in **Supplementary Table 16**. All mutations identified in tumors were confirmed by independent PCR and Sanger sequencing in the specific tumors and their paired normal tissue.

Statistical analysis of clinico-pathological correlations. Comparison of the NS/S ratio with clinico-pathological parameters was performed using the Mann-Whitney *U* test. Analysis of ARID1A mutation or alteration frequency with clinico-pathological parameters was performed using either the chi-squared test or Fisher's exact test, when appropriate. Sixty-two individuals with gastric cancer who had undergone curative resection were analyzed for factors related to disease-free survival. Sex, age, tumor stage, site and differentiation, Lauren's tumor type, *H. pylori* infection, MSI status, EBV status, *TP53* mutation and ARID1A alterations were analyzed for their prognostic value using Kaplan-Meier survival analysis with log-ranked tests as well as Univariate Cox regression analysis. Significant factors in Univariate analysis (tumor stage, Lauren's tumor type, MSI status and ARID1A alterations) were further subjected to a multivariate Cox regression analysis in a forward stepwise manner.

Mononucleotide repeat mutation frequency analysis in MSI gastric cancers. The MSI cancers tend to have frequent random mutations involving mononucleotide repeats, and the mutation frequency increases with the length of the mononucleotide repeat. True driver genes targeted by the MSI mechanism generally show a substantially elevated mutation rate in their coding repeats compared to the background mutation rate for the same repeat length[50]. Thus, to address whether the observed high mutation rate of *ARID1A* in MSI gastric cancers is simply due to a high background rate of mutations in these tumors, we first estimated the background mutation rate for coding mononucleotide repeats in relation to repeat length across the entire exome for the four MSI gastric cancers in the discovery cohort (for which whole-exome sequencing data were available). Specifically, for each mononucleotide repeat length we

doi:10.1038/ng.982

calculated the total number of coding repeats mutated in the four MSI gastric cancers and determined the total number of coding repeats that were adequately covered among all targeted regions by SureSelect (including 16,568 genes). Mutation frequency for each mononucleotide repeat in *ARID1A* in all 23 MSI gastric cancers was then compared to the global background mutation rate using the chi-squared test (**Supplementary Table 15**).

Driver gene analysis. Please refer to **Supplementary Note** and **Supplementary Table 17**.

49. Yuen, S.T. *et al.* Similarity of the phenotypic patterns associated with BRAF and KRAS mutations in colorectal neoplasia. *Cancer Res.* **62**, 6451–6455 (2002).
50. Woerner, S.M. *et al.* Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes. *Oncogene* **22**, 2226–2235 (2003).