Global and disease-associated genetic variation in the human Fanconi anemia gene family

Kai J. Rogers1,†, Wenqing Fu2,†, Joshua M. Akey2 and Raymond J. Monnat Jr2,3,*

1Department of Microbiology, 2Department of Genome Sciences and 3Department of Pathology, University of Washington, Seattle, WA 98195, USA

Received April 8, 2014; Revised June 9, 2014; Accepted July 30, 2014

Fanconi anemia (FA) is a human recessive genetic disease resulting from inactivating mutations in any of 16 FANC (Fanconi) genes. Individuals with FA are at high risk of developmental abnormalities, early bone marrow failure and leukemia. These are followed in the second and subsequent decades by a very high risk of carcinomas of the head and neck and anogenital region, and a small continuing risk of leukemia. In order to characterize base pair-level disease-associated (DA) and population genetic variation in FANC genes and the segregation of this variation in the human population, we identified 2948 unique FANC gene variants including 493 FA DA variants across 57 240 potential base pair variation sites in the 16 FANC genes. We then analyzed the segregation of this variation in the 7578 subjects included in the Exome Sequencing Project (ESP) and the 1000 Genomes Project (1KGP). There was a remarkably high frequency of FA DA variants in ESP/1KGP subjects: at least 1 FA DA variant was identified in 78.5% (5950 of 7578) individuals included in these two studies. Six widely used functional prediction algorithms correctly identified only a third of the known, DA FANC missense variants. We also identified FA DA variants that may be good candidates for different types of mutation-specific therapies. Our results demonstrate the power of direct DNA sequencing to detect, estimate the frequency of and follow the segregation of deleterious genetic variation in human populations.

INTRODUCTION

Fanconi anemia (FA) is an uncommon recessive genetic disease that results from loss of function of any of the 16 genes belonging to the FANC gene family. FA is characterized by an elevated risk of developmental abnormalities, progressive anemia leading to bone marrow failure and an elevated risk of leukemia in the first decade of life (1–4). Bone marrow transplantation can provide an effective cure for marrow failure in FA patients, but does not address associated developmental defects or attenuate the risk of developing carcinomas of the head and neck and anogenital region or the small continuing risk of leukemia (3–9). The proteins encoded by the 16 FANC genes function together in a common DNA damage recognition and repair pathway with other proteins (10–12) (Fig. 1). Of the 16 FANC proteins, 6 were previously identified as part of different DNA damage repair pathways that may interact with or complement the FA pathway: FANCD1/BRCA2, FANCJ/BRIP1, FANCN/PALB2, FANCO/RAD51C, FANCP/SLX4 and FANCQ/XPF/ERCC4 (10–12).

We took advantage of the molecular characterization of FANC variation in FA patients, together with genomic sequencing of the FANC genes as part of the 1000 Genomes Project (1KGP) (13) and the Exome Sequencing Project (ESP) (14,15) to systematically compare all base pair-level, disease-associated (DA) FANC gene variation with variation identified in the 7578 individuals included in these sequencing projects. Our goals were to determine how the spectrum (types and sites) of FANC gene variation differed between individuals affected with FA versus population-derived variation data; to identify new, potentially deleterious FANC variants in population sequencing data; to quantify how well functional prediction algorithms correctly identified known, DA FANC missense variants and to provide an estimate of the carrier frequency of deleterious FANC variants by direct variant counting in population data. A subsidiary goal was to identify subsets of

†The first two authors should be regarded as joint first authors.
*To whom correspondence should be addressed. Tel: +1 2066167392; fax: +1 2065433967; Email: monnat@u.washington.edu

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
FA DA variants that may be good candidates for different types of mutation-specific therapies.

RESULTS

Molecular spectrum of FANC variation

We identified 2948 unique small base pair-level variants among the 16 FANC genes. These included 493 DA or disease-causative (DC) variants listed in the Fanconi Anemia Mutation Database (FAMutDB), and 2508 genomic variants from ESP and 1KGP data (Table 1). We compiled and maintained two separate databases, one containing only unique variants, and a second that included data on variant frequency. For the following analyses we used only unique variants. We identified 1753 additional unique, base pair-level variants in FANCD1/BRCA2 using data in the Breast Cancer Information Core (BIC) (16). These variants showed extensive overlap in other mutation databases, and a majority had no defined association with FA. Thus, we excluded BIC-derived variants from subsequent analyses apart from one comparison mentioned below. BIC data were accessed on 30 May 2013.

The molecular type distribution of FANC variants was heavily biased in favor of base substitutions: 2689 (91.2%) single base substitutions were observed, together with 114 (3.9%) single base insertions/deletions, 139 (4.7%) simple 1 bp variants and 6 (0.2%) complex 1 bp variants (Table 1). In ESP/1KGP data, we found 97.4% of all FANC variants were single base substitutions.

![Figure 1. Model of the Fanconi functional pathway and protein components.](https://academic.oup.com/hmg/article-abstract/23/25/6815/571236)

**Table 1.** FANC gene variants by type in FAMutdB and ESP/1KGP sequencing data

<table>
<thead>
<tr>
<th>Gene</th>
<th>FAMutdB (n = 493)</th>
<th>ESP/1KGP (n = 2508)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 bp subs</td>
<td>± 1 bp indel</td>
</tr>
<tr>
<td>FANCL</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FANCD2</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>FANC</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>FANCG</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>FANC</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>FANC</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>BRCA2</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>FANC</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>FANCI</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>FANC</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>ERCC4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>PALB2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>FANCA</td>
<td>154</td>
<td>41</td>
</tr>
<tr>
<td>RAD51C</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>BRIP1</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>FANC</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

Total (%) 295 (59.8%) 86 (17.4%) 106 (21.5%) 6 (1.2%) 2443 (97.4%) 32 (1.3%) 33 (1.3%) 0 (0)

*a FANC genes are ordered according to their locations in the human genome beginning with chromosome 1. Several have well-known common names, e.g. BRCA2 = FANCD1; ERCC4 = XPF = FANCP; PALB2 = FANCN; RAD51C = FANCO and BRIP1 = BACH1 = FANCJ.

**b** DA genetic variation was downloaded from the FAMutdB (http://www.rockefeller.edu/fanconi/). Human genomic variation data were obtained from the 1KGP (http://www.1000genomes.org) and ESP data archives. Single base substitutions in ESP were identified in Fu et al. (9), whereas insertion/deletion and >1 bp variant data were downloaded from the Exome Variant Server (http://evs.gs.washington.edu/EVS/). The full data set detailing the above variants by FANC gene and source is available upon request.
substitutions. In contrast, 59.8% of variants in the DA FAMutdB were base substitutions. This skewing likely reflects the strong enrichment for variants in the FAMutdB data that lead to loss of function. The molecular type distribution of variation in the X-linked FANCB gene did not differ from the other 15 autosomal FANC genes (Fisher's exact test, $P = 0.20$ and $0.82$ for data from FAMutdB and ESP/1KGP, respectively). Figure 2 provides a visual representation of the broad distribution of base pair-level variation by type across the open reading frames of the three FANC genes mutated in a majority of FA patients: FANCA, FANCC and FANCG.

We compared variants in FANCD1/BRCA2 using data from the FAMutdB and from the BIC in order to determine whether there were distinguishing molecular characteristics of variation that was FA-associated as opposed to breast- or ovarian cancer-associated. Inherited deleterious variants in both copies of the FANCD1/BRCA2 gene give rise to a strongly penetrant early-onset FA phenotype dominated by features of FA in conjunction with a high risk of leukemia, brain tumors (most often medulloblastoma) and Wilms’ tumor. This is in contrast to single inherited deleterious FANCD1/BRCA2 variants that confer a strongly elevated lifetime risk of breast and ovarian cancer (17–22). FA-associated FANCD1/BRCA2 gene variants were distributed similarly to those found in the BIC and in the same proportions, with a small difference in the fraction falling in exon 11 in FA patients (40%, $n = 14$) versus ESP/1K individuals (50%, $n = 196$). Of note, only two variants found in the FAMutdB were not already present in BIC data (results not shown). Thus, FA-associated FANCD1/BRCA2 variants did not differ significantly in location, type or predicted severity from those identified in breast/ovarian cancer patients. Rather, it is the inheritance of two deleterious FANCD1/BRCA2 alleles, which confers a strongly penetrant, early FA phenotype.

Figure 2. Landscape of base pair-level genetic variation in FANCA/C/G. (A) Biallelic deleterious variants in FANCA, FANCC or FANCG have been identified in 67% of the FA patients included in FAMutdB data (see Table 1). The open reading frame of each gene is represented by the central rectangle, where vertical lines indicate exon boundaries. Amino acid residue and nucleotide counts for open reading frames are indicated, respectively, above and below the horizontal line. The locations and molecular variant type of DA (filled symbols) and ESP/1KGP sequencing project genomic variants (open symbols) are indicated by circles (base substitutions), upright or inverted triangles (indels) or diamonds (for more complex mutations; see text for additional detail). (B) Variant landscape key for symbols indicating different molecular types of FA DA variation (filled symbols) and ESP/1KGP genomic variants by molecular type (open symbols).
Mutability and mutation hotspot analysis

*FANCA, FANCC* and *FANCG* are the most commonly mutated genes in FA patients, accounting for 67.7% of unique variants contained in the FAMutdB. These three genes constitute 14.6% of the total open reading frame and splice junction sequences that define the 16 *FANC* genes. We thus asked whether these genes were disproportionately mutable or harbored mutation hotspots. In order to do this we used a recently developed probabilistic model (23) to confirm an unusually high proportion of FA DA variants concentrated in these three genes ($\chi^2$-test, $P = 1.05 \times 10^{-160}$ for *FANCA*, $P = 2.43 \times 10^{-11}$ for *FANCC* and $P = 3.73 \times 10^{-17}$ for *FANCG*) (Fig. 3A) despite their containing only 17.5% of genomic variation among the *FANC* genes in a combination of ESP and 1KGP data. Thus despite their importance in FA, *FANCA, FANCC* and *FANCG* do not appear to be intrinsically more mutation-prone than the other 13 *FANC* genes (Fig. 3B).

We also searched for mutation hotspots in the *FANC* genes, beginning with genomic data and using a probabilistic model. We defined mutation hotspots using a 5bp window to identify gene regions that were significantly enriched for variation compared with the expected ($\chi^2$-test, $P < 0.05$). The rationale for using a 5 bp window was that windows <5 bp might miss various types of known hotspots consisting of mononucleotide runs or repeats, whereas much larger windows of, e.g. 20 or 50 bp, might reduce the sensitivity of hotspot detection by averaging signal over a larger number of nucleotides. A total of 858

![Figure 3](https://academic.oup.com/hmg/article-abstract/23/25/6815/571236)

**Figure 3.** *FANC* gene mutation hotspot, gene and type-specific analyses. (A) Mutation hotspots containing significantly more mutations than predicted were identified in all 16 *FANC* genes using ESP/1KGP variant data and a mutation hotspot window of 5 bp (see text for additional detail). We identified 858 potential mutation hotspots using genomic data for each gene, listed left to right along a horizontal axis in their chromosomal location order and color-coded. The footnote in Table 1 cross-correlates *FANC* gene designations with alternative gene names. The location of each hotspot (identified by a $\chi^2$-test $P < 0.05$ threshold depicted by the dotted line) is indicated by gene-specific, color-coded dots with hotspot significance indicated by the $-\log_{10}(P)$ value. FA DA variants in each hotspot window were then plotted below the horizontal line with the number of different variants/hotspot indicated. No significant difference was observed for FA DA variants located in or outside of mutation hotspots (Mann-Whitney rank sum test $P = 0.44$). (B) The comparison of the observed variants in ESP/1KGP and FAMutdB with the number expected for each gene, respectively, by the $\chi^2$-test. Significant results ($P < 0.001$) are marked by an asterisk (**), with significantly overrepresented results in red and underrepresented in blue. *FANC* genes are again ordered according to their locations in the human genome. (C) Comparison of the number of observed and expected variants in ESP/1KGP (above) and FAMutdB variants (below the horizontal axis) as a function of predicted functional consequence (i.e. synonymous, missense, nonsense, splice and frameshift). The $\chi^2$-test was used to compare the observed number with the expected number. Significant results ($P < 0.001$) are marked by an asterisk (**), with significantly overrepresented results in red and underrepresented in blue.
putative mutation hotspots were identified by these criteria in ESP/1KGP data (Fig. 3A). Nearly half (46.2%) of population variants were located in these mutation hotspots despite hotspots representing only 7.4% of the total mutable sequence, a hotspot enrichment for variation of 6.24-fold.

In order to gain additional insight into the mechanistic origins of FANC variation, we determined whether sequence features of hotspot regions might be promoting variation or the generation of specific molecular types of variant. This analysis was performed by looking for the overrepresentation of known mutable sequences including tandem repeats (identified by Tandem Repeat Finder) and methylated cytosine-guanine (CpG) dinucleotides (identified across 82 cell types from ENCODE) within or bridging hotspot boundaries. CpG sites were slightly enriched in mutation hotspots using a hotspot window of 5 bp, compared with non-hotspots [Fisher’s exact test, \( P = 0.09, \text{odds ratio (OR)} 1.562 (0.952–2.562) \)]. In contrast, methylated CpG sites were significantly enriched in mutation hotspots [Fisher’s exact test, \( P = 0.0009, \text{OR} 2.945 (1.639–5.292) \)]. When we defined mutation hotspots by using a window of size 10 or 20 bp, methylated CpG sites were hotspot-enriched, though not significantly [Fisher’s exact test, \( P > 0.05, \text{OR} 1.649 (0.709–3.835) \) for 10 bp hotspots, and \( \text{OR} = 1.321 (0.472–3.694) \) for 20 bp hotspots]. We observed no significant association of hotspots with tandem repeats using hotspot windows of 5, 10 or 20 bp (Fisher’s exact test, \( P > 0.05 \)). FA DA variants identified in the FAMutdB were not enriched in mutation hotspots (5 bp window) when compared with non-hotspots (Mann–Whitney test, \( P = 0.44 \)). This lack of significant enrichment was also observed when using hotspot windows of 10 or 20 bp (Mann–Whitney test, \( P > 0.05 \)).

Visual inspection of the DNA sequence of the 16 ‘hottest’ FANC mutation hotspots, as defined by their \(-\log_{10}(P)\) value (Fig. 3A), was also done to search for additional clues to mutation mechanism. These hotspots were located in seven different FANC genes and contained 56 different variants. All hotspot-associated variants were in an 11 bp window that contained the initially identified 5 bp hotspot sequence and flanking DNA defined by the 5′-most and 3′-most mutated base pairs across all 16 hotspots. Hotspot variants were predominantly (93%) base substitutions, with a strong bias in favor of transitions over transversions (39 transitions versus 13 transversions). We observed 2 single base deletions at these 16 hotspots, together with 1 and 10 bp insertions. The only conspicuous DNA sequence motif found in 4 of the hotspot-associated 11 bp windows was mononucleotide runs of 4, 5 or 7 G:C base pairs that themselves contained from 1 to 4 variants/hotspot. Several of the hotspot base substitutions and one of the single base deletions may thus have arisen by DNA strand slippage with or without base insertion during DNA replication.

Previous studies have observed an increase in base miscalls and other errors in GC-rich regions sequenced using next-generation sequencing (NGS) technologies (24). Thus we determined whether 91 reported motifs that induce NGS sequencing errors (25) were present or enriched in our hotspot-associated 11 bp windows. However, none of these 91 motifs was identified in any of our 16 hottest mutation hotspot sequences. There was a modest enrichment of error-inducing sequence motifs across all mutation hotspots [Fisher’s exact test, \( P = 0.03, \text{OR} 1.87 (1.08–3.23) \)]. Thus variations in ESP/1K data that fell in our 16 hottest mutation hotspots are unlikely to be sequencing miscalls. Of note, the subset of FA-associated variants identified in these hotspots were originally identified and confirmed by capillary, as opposed to NGS, sequencing.

Predicted variant functional consequences

The distribution of variants by predicted consequence in ESP/1KGP data was 66.0% missense and 29.1% synonymous amino acid substitutions, together with 4.9% for three other functional categories (nonsense, splice-altering changes and insertion/deletions causing a frameshift with premature chain termination). The corresponding distribution for FA DA variation was 33.3% missense, 2.6% synonymous, 21.9% nonsense, 2.0% splice-altering and 40.2% frameshifts. Variants that truncated a FANC protein (e.g. variants leading to frameshifts and chain termination) were significantly enriched in the FAMutdB database (Fig. 3C). We found 53 (10.7%) FA-associates variants in ESP/1KGP population sequencing data: 49 in ESP, and 28 in 1KGP data. There were six nucleotide positions at which we observed different polymorphisms in both FAMutdB and in ESP/1KGP data that led to the same predicted protein changes. There were also 14 different nucleotide positions at which we observed both disease-associated FAMutdB and ESP/1KGP variants. These positions in FAMutdB data contained nine variants predicted to lead to one missense substitution, four nonsense codons leading to chain termination, and nine frameshifts leading to chain termination. In ESP/1KGP data, variation at these positions was predicted to generate 11 different missense and 3 synonymous amino acid substitutions. A closer look at insertion–deletion (indel) variants revealed that 99% caused disruption of the open reading frame either by a direct reading frame shift or by splice disruption; <1% resulted in the addition or removal of a codon while preserving the open reading frame. Thus virtually all of these FA-associated variants are predicted to disrupt FANC protein structure and/or function.

Finally, the predicted consequence of variants identified in the X-linked FANC A gene did not differ from variants observed in the other 15 autosomal FANC genes (Fisher’s exact test, \( P = 0.52 \) and 0.17 for data from FAMutdB and ESP/1KGP, respectively).

In order to determine what fraction of all possible FANC gene single base variants were likely to be deleterious, we first generated a list of all possible single base variants across the 16 FANC genes and then imputed the subset giving rise to 123 073 potential missense substitutions. We then determined whether these potential missense substitutions were likely to be deleterious based on a majority rule using six different, widely used algorithms including SIFT (26,27), PolyPhen2 (28), a likelihood ratio test (LRT) (29), Mutation Taster (30), GERP++ (31) and PhyloP (32). Missense substitutions were considered deleterious if called by four or more of the six different prediction algorithms. Using this approach, 26.9% (33 148 of 123 073) potential missense substitutions were annotated as deleterious. We used the same approach to annotate all of our FA DA or genomic variants leading to missense substitutions. As expected, FA-associated variants leading to missense substitutions were more frequently predicted to be highly deleterious, 43.3% (71/164) as compared with ESP/1K missense variants, 23.6% (391/1,655) (Fig. 4).
Potential for mutation-specific functional rescue

A subset of FA DA variants might be candidates for therapeutic exon skipping, or stop codon read-through to restore the open reading frame (33–39). In order to assess mutation-specific therapeutic potential, we determined the proportion of ‘skippable’ exons in the 16 FANC genes and the number of chain-terminating nonsense codons that were well positioned to facilitate premature termination codon (PTC) suppression or ‘read-through’. The 16 FANC genes collectively consist of 307 exons, 104 of which (34%) could be skipped without disrupting the downstream open reading frame. These individually ‘skippable’ FANC exons collectively contained 32% of all reported FA DA variants. We identified 2702 base substitutions in FAMutdB and ESP/1KGP data, of which 144 (5.3%) created a new nonsense codon within a FANC gene open reading frame. Among FA DA variants, 108 base substitutions (21.9%) created a new nonsense codon. Ten of these (9.3% of all DA variants) were located within 50 nucleotides of the terminal intron–exon junction and thus might be good candidates for PTC suppression or read-through (36–39).

Population-based variant and pathway analysis

In order to determine the prevalence of potential disease-causing FANC variants in data from the ESP/1KGP projects, we used two sets of FA-associates: all FA DA variants in the FAMutdB except for a single, high frequency polymorphic variant; and a more stringently defined variant subset with a very high likelihood of being deleterious by virtue of generating null alleles (disease-causing or DC variants). Among 493 variants identified in the FAMutdB, 387 were predicted to be disease-causing (78.5%). Of note, among the clinically ascertained DA variants also identified in ESP/1KGP data, only 39.6% (21/53) were predicted to be putatively DC. In order to examine the segregation of these DA and DC FANC variants in ESP and 1KGP data, we determined the number of individuals who carried zero, one or more than one DA or DC variant in one or more FANC genes. We also searched for potential FA patients, defined by their carrying two deleterious variants in a single FANC gene. The proportion of ESP/1KGP individuals carrying single DA alleles was remarkably high: 78.5% (5950/7578; Table 2A). These frequency estimates were 18.4-fold higher than the frequency of individuals carrying the more restrictive subset of FA DC variants.

In ESP data, 8.9% of African Americans carried one DC allele. We also identified three potential FA patients who carried the same FANCC variant allele in homozygous form at position chr9: 97 934 359 (Table 2B). This variant had a derived allele frequency (DAF) of 0.045 in African American ESP data and was observed in heterozygous form in European American ESP data (two occurrences, DAF = 0.0002), in African 1KGP data (18 occurrences, DAF = 0.036) and once in American 1KGP data (DAF = 0.003). Among European Americans in ESP data, 2.1% carried one DC allele, although no homozygous individual was identified (Table 2). Among 1KGP data, seven DC variants were identified in 36 individuals: 18 African-ancestry individuals, 15 European-ancestry individuals and 3 American-ancestry individuals; no homozygous potential FA patients were identified. In aggregate, 4.3% (324 of 7578) of all individuals represented in the ESP/1KGP populations carried at least one DC allele, and we identified three potential FA patients in the combined ESP/1KGP population, a frequency of 4 × 10⁻⁴ (3 of 7578). These estimates of the frequency of individuals carrying ≥ 1 FA DA or DC variant are substantially higher than previous estimates of carrier frequency that have ranged from 0.55 to 1.08%, depending upon the reference population and method used for estimation (2,40).

We also asked how prevalent partial loss of FA pathway function might be by searching for DA or DC alleles in two or more FANC genes in single individuals. Among ESP/1KGP individuals, 29.6% (2242/7578) carried a DA variant in >1 FANC gene, though none carried ≥2 DC variants in different FANC genes (Table 2). These predicted frequencies did not differ from expectation (Table 2). Thus partial loss of FA pathway function (Fig. 1), driven by deleterious variation in ≥2 FANC genes, does not appear to be subject to negative selection.

DISCUSSION

This analysis focused on characterizing base pair-level genetic variation in the 16 genes of the FA gene family (FANC). We used the well-documented, clinically ascertained collection of 493 unique FANC variants from the FAMutdB (41), together with genetic variation data in the 16 FANC genes identified in 7578 individuals in the ESP and 1KGP genomic sequencing projects (13,15,42). These sources together included 493 variants ascertained in FA patients or family members, and 2508 variants identified in individuals included in the ESP and 1KGP genomic sequencing projects. Insertions and deletions leading to frameshifts predominated in FA patients, whereas single base substitutions represented the majority of variants (97.4%) in ESP/1KGP data (Table 1). Several of the FANC genes have a substantial proportion of larger variants including deletions and duplications identified in FA patients (43,44). It should be possible to systematically compare the FA DA large-scale variation with...
Table 2. Distribution of FA disease variants across the 16 FANC genes in ESP/1KGP populations

<table>
<thead>
<tr>
<th>Variants/genes*</th>
<th>ESP_AA</th>
<th>ESP_EA</th>
<th>1KGP_AFR</th>
<th>1KGP_EUR</th>
<th>1KGP_ASI</th>
<th>1KGP_AME</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA variants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0,0</td>
<td>209</td>
<td>211.3</td>
<td>0.87</td>
<td>1185</td>
<td>1192.3</td>
<td>0.67</td>
</tr>
<tr>
<td>S,S</td>
<td>631</td>
<td>617.9</td>
<td>0.53</td>
<td>2141</td>
<td>2073.4</td>
<td>0.03</td>
</tr>
<tr>
<td>M,S</td>
<td>222</td>
<td>220.1</td>
<td>0.89</td>
<td>216</td>
<td>304.7</td>
<td>10^{-5}</td>
</tr>
<tr>
<td>M,M</td>
<td>1155</td>
<td>1167.7</td>
<td>0.59</td>
<td>756</td>
<td>727.6</td>
<td>0.25</td>
</tr>
<tr>
<td>DC variants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0,0</td>
<td>2019</td>
<td>2024.1</td>
<td>0.70</td>
<td>4208</td>
<td>4207.3</td>
<td>0.90</td>
</tr>
<tr>
<td>S,S</td>
<td>198</td>
<td>192.7</td>
<td>0.69</td>
<td>90</td>
<td>89.8</td>
<td>0.99</td>
</tr>
<tr>
<td>M,S</td>
<td>0</td>
<td>0.0</td>
<td>NA</td>
<td>0</td>
<td>0.0</td>
<td>NA</td>
</tr>
<tr>
<td>M,M</td>
<td>0</td>
<td>0.2</td>
<td>0.64</td>
<td>0</td>
<td>0.9</td>
<td>0.35</td>
</tr>
</tbody>
</table>

(A) Individuals who carry at least one derived allele DA or DC variant

The observed (Obs.) and expected (Exp.) number of individuals who carry none, one or more than one DA or DC variant were categorized into four groups: (0,0), (S, S), (M,S) and (M,M) individuals carried, respectively, no variants in any FANC gene; 1 variant in one gene; >1 variant in one gene or >1 variant in multiple genes, respectively. The observed distribution was compared with the expected one from 10 000 simulations, with significantly marked in bold ($P < 0.05$).

*This analysis was conducted using 52 DA variants excluding one major derived allele, or a more restrictive set of 21 DC variants (see text for additional detail).
population variation data once larger numbers of these large scale DA and population variants are defined at the molecular level (45).

We found 10.7% of clinically ascertained FA DA variants in population sequencing data (49 variants in ESP, and 28 variants in 1KGP data), together with a surprisingly high proportion of ESP/1KGP individuals who carried a known FA DA variant (78.5%, or 5950/7578; Table 2A). This is a substantially higher frequency for deleterious genetic variation in the FANC gene family that suggested by previous estimates of carrier frequency developed from FA disease prevalence data (2,40). This seeming discrepancy is likely accounted for by our use of a sequence-based approach, which enumerated variation across all 16 FANC genes and included all documented base pair-level DA variants. Neither of these variables is typically fully accounted for in carrier frequency estimates based on disease prevalence. Thus our results are in accord with and extend recent analyses showing that individuals may carry 100s of potentially deleterious variants (46,47). No readily identifiable heterozygote phenotype has been identified for DA FANC variants (2–4,7) (however, see (48) for one intriguing clue). The high proportion of individuals in ESP/1K data that carry multiple DA FANC variants in ≥2 FANC genes raises the possibility of pathway functional haploinsufficiency. This might be clinically penetrant in specific genetic backgrounds or in response to unusual environmental exposure to, e.g. aldehydes or DNA cross-linking drugs or agents.

Despite the high carrier frequency for DA FANC variants, we identified only three potential FA patients in ESP/1KGP populations or a frequency of 4 × 10−4 (3 of 7578). This is likely to underestimate of the frequency of potential FA patients, as we used only the most stringently defined set of FA DC variants and our genomic data were not haplotype resolved. The FA clinical phenotype shows a wide range of penetrance, with an increasing number of individuals being recognized to have FA only in adulthood on the basis of hematologic complications or neoplasia (2–4). These ‘atypical’ FA patients may be of high interest to examine further to identify additional genetic or environmental modifiers of FA disease penetrance.

When we determined how well functional prediction algorithms correctly identified known, FA DA missense variants, even a consensus approach using six widely employed functional prediction algorithms correctly identified fewer than half of FA DA variants. Fully one-third of FANC DA variants had low or no ‘predicted deleterious’ scores (scores of 0–2; Fig. 4). These results emphasize the continuing need to couple computational and experimental approaches to determine the functional consequences of potential DA variants and variants of unknown functional significance (VUSs).

Our analysis identified a large subset of FA-associated variants that are good candidates for mutation-specific therapies. Fully a third (104, or 34%) of all FANC exons are potentially ‘skippable’, and these exons contain 32% of all FA DA variants. Exon skipping as a therapeutic strategy has been most extensively explored in Duchenne muscular dystrophy (33–35). Among DA variants, 108 (21.9%) create new nonsense codons. Ten of these (9.3%) are within 50 bp of the terminal intron–exon junction, and thus good candidates for PTC suppression or read-through therapies (36–39); efficacy depends on stop codon identity, location and sequence context (49). Two prevalent FANC DA variants may be especially good candidates for PTC therapy: both create new stop codons in the terminal exon of FANCC and one, generated by a FANCC c.1897 C>T base substitution, is present in 17% of the FANCC complementation group (or ~2% of all) FA patients (41). FA DA missense variants might also be good candidates for functional rescue or complementation by drugs or small molecules. This strategy, which depends on the biochemical consequences of specific variants, has been successful in another recessive disease, cystic fibrosis (50–57).

In summary, we have defined base pair-level genetic variation including DA variation in FA patients and in 7578 individuals included in the ESP and 1KGP genomic sequencing projects at 57 240 potential mutation sites across all 16 FANC genes. We identified a remarkably high frequency of FA DA variants in population data despite the rarity of FA, have provided clues to the mechanistic origins of FANC genetic variation and have identified the subset of FA DA variants that may be amenable to mutation-specific therapies. Our analytical approach is general and thus can be used to gain insight into the nature and functional consequences of genetic variation in other recessive DNA damage response/repair syndromes as well as in many other single gene or multi-gene disorders.

MATERIALS AND METHODS

Variation data sources

The principal source for FANC gene DA variation was the FAMutDB (http://www.rockefeller.edu/fanconi/) maintained by Dr Arleen Auerbach and colleagues at the Rockefeller University (41). The FAMutDB catalogs FA patient-associated genetic variation, and the number of times each unique variant has been reported in the context of FA. Human genomic variation data were obtained from 1KGP (13) and ESP (14,15) data archives. 1KGP Phase I data include 1092 individuals sampled from 14 populations drawn from Europe, East Asia, sub-Saharan Africa and America. This data set aimed to obtain genetic variation data that would be broadly representative for the vast majority of individuals within a continent, though with no associated, publicly available phenotype or clinical information (13). ESP data were generated as part of the NHLBI-funded GO (grand opportunity) ESP. ESP samples were drawn from several well-phenotyped populations (i.e. 4298 European Americans and 2217 African Americans) and sequenced with the goal of discovering new genes and mechanisms contributing to heart, lung and blood disorders (15). All study participants provided written informed consent for the use of their DNA in studies aimed at identifying genetic risk variants for disease and for broad data sharing.

Additional data on the FANC1/BRCA2 gene were downloaded from the NHGRI Breast and Ovarian Cancer Mutation Database (BIC) (16). We also included additional variation data identified by PubMed and Google Scholar searches that were not captured in the FAMutDB. The final data capture using database sources and publication searches was completed on 1 November 2013.

Molecular spectrum of FANC variation

Variants were categorized into four molecular types: single base substitutions; single base insertions or deletions; and variants
that involved >1 bp and were either ‘simple’ (e.g. insertion or deletion of multiple base pairs with no other changes) or ‘complex’ (e.g. combinations of base pair changes, insertions, deletions or other rearrangements). The reference DNA sequence, open reading frame and inferred protein sequence for each FANC gene were taken from the most recent RefSeq gene, mRNA and protein accessions for each gene. The positions of amino acid residues were annotated according to the largest isoform for each gene, and genome sequences from human genome assembly GRCh37.p13 were accessed via the Locus Reference Genomic (LRG) initiative website (http://www.lrg-sequence.org/) to provide stable, curated reference sequences. Sequence variation was cataloged using Human Genome Variation Society nomenclature guidelines (58,59). All unique DA variants and their locations within the 16 FANC genes were compiled, together with a list of the number of individual reports of each DA variant to provide a measure of variant prevalence in FAMutdB data. An analysis of global genomic variation in the 16 FANC genes was performed by identifying and categorizing all small genetic variants identified in 1KGP and ESP data using the same reference sequences and mutation nomenclature guidelines. Our analysis focused on coding region variants with the exception of those occurring within canonical splice sites, as we reasoned that a majority of DA variation was within coding regions. We excluded large structural rearrangements such as duplications, deletions or rearrangements involving ≥ 1 exon from subsequent analyses, and note that when applied to FANCA, the most frequently mutated gene in FA patients, this led to the exclusion of ≤ 10% of the DA variants cataloged in the FAMutdB. We did not systematically analyze FANC gene copy number variants. Our final data set of analyses included 2948 unique FANC gene variants, across 57 240 potential base pair sites, in the 16 FANC human genes.

Mutation hotspot analysis

We used a previously developed probabilistic model (23) to determine whether mutation ‘hotspots’ existed in any of the FANC genes. This model incorporates an overall rate of mutation and relative locus-specific rates, based on fixed differences between humans and chimpanzees, for each gene coding and splice junction sequence. It also takes into account codon structure, and transition/transversion ratios in order to predict the relative probability of observing mutations in a specific gene region. This model was applied by taking a total of \( N_T \) variants observed in FANC genes, corresponding to \( N_T \) mutation events assuming an infinite alleles model (60), and randomly distributing these mutation events according to their relative probabilities to obtain the expected variant distribution in the absence of hotspots. Then we compared the observed number of variants mutations identified in ESP/1KGP data with the number expected by use of a \( \chi^2 \)-test. We looked at hotspots defined by a small sliding window (e.g. 5 bp) within each gene, as well as for each gene as a whole, prior to examining the nature of variants at these hotspots in FAMutdB and ESP/1KGP data.

In order to determine whether there might be an underlying sequence-dependent basis for mutations or mutation hotspots, we downloaded the coordinates of CpG sites and their methylation status using ENCODE project data, where reduced representation bisulfite sequencing (RRBS) was used to quantitatively profile DNA methylation at an average of 1.2 million CpGs in each of 82 cell lines and tissues (63). We defined a CpG site as ‘methylated’ when the site was methylated in >50% of molecules sequenced on average. Fisher’s exact test was then used to compare the distribution of simple tandem repeats or methylated CpG sites across mutation hotspots and non-hotspots. We also visually examined the sequences of the 16 most significant hotspots and their associated variants together with the flanking 20 bp up- and downstream of each hotspot for mechanistic clues to mutagenesis (see ‘Results’ for additional description).

Variation functional consequences

In order to assess variant functional consequences, we first predicted the consequences of all possible single base changes by determining the number and proportion of all single base changes at each base pair position that did not modify the amino acid sequence (synonymous substitutions), or that led to an amino acid substitution (missense substitutions), a stop/sense codon, or altered the reading frame by inserting or deleting base pairs or by leading to a splice alteration. In order to identify missense substitutions that were potentially deleterious, we used six different functional prediction methods in concert together with a majority rule (four or more of the six ‘predicting deleterious’) as in previous studies (15). The functional prediction algorithms used were SIFT (26,27), PolyPhen2 (28), a LRT (29), Mutation Taster (30), GERP++ (31) and PhyloP (32). Functional predictions were made for all FA DA variants; for all variants identified in ESP/1KGP sequencing data and for all possible single base substitutions (\( n = 123 073 \)) predicted to lead to missense substitutions across all 16 FANC genes.

Potential of disease-associated variants for functional recovery

In order to determine what fraction of DA FANC variants might be candidates for functional recovery by exon skipping to restore the open reading frame, we assessed all 16 FANC genes using the FAMutdB Reading Frame Checker utility (41) to determine whether or not an exon skip would result in a reading frame shift. We then determined what fraction of FA DA variants were candidates for the skipping of a single exon to restore the protein open reading frame. For simplicity and with an eye to potential translation, we investigated only instances in which the skipping of single exons removed a DA variant while restoring the open reading frame. The potential of PTC suppression or ‘read-through’ was assessed by identifying FA DA PTCs that were ≤50 nucleotides upstream of the last exon—intron junction, as these PTCs are the most likely to be rescued by PTC read-through as they often escape nonsense-mediated decay (36–39,49).

A population-based, pathway-focused analysis of FANC variation

In order to determine the distribution of FA DA variants in sequencing data, we first determined the frequency of DA FA
variants in the 4298 European Americans and 2217 African Americans included in ESP data and in the 1063 unrelated individuals (excluding related individuals with second order) with ancestry from four geographical regions (Europe, East Asia, sub-Saharan Africa and America) included in 1KGP data (13,15,42). FA DA variants (DA variants) were defined by their presence in the FAMutDB and their low frequency. We excluded one derived allele in FANCA (c.2346 C>T) with a very high allele frequency (0.6) that was likely fortuitously captured in FAMutDB data simply by virtue of prevalence rather than disease association. In order to avoid the uncertainty of phasing, especially for rare variants, we used only genotyping data.

We next used DA FANC variants to identify potential FA patients in the 1KGP/ESP populations and determine the frequency of DA alleles to compare with previous estimates of carrier frequency from demographic data (40); and whether individuals with multiple variants in the same or in different FANC genes were observed at the expected frequency in the 1KGP/ESP cohorts. The last of these analyses was undertaken to determine whether DA variants in more than one FANC gene might be subject to negative selection. Two complimentary analyses were performed: one in which all DA variants defined by their presence in the FAMutDB were included; and a second, more stringent analysis that included only the subset of DA variants (here termed DC variants) that were highly deleterious: nonsense, splice-altering and frameshift-promoting variants, together with the subset of missense variants predicted to be highly deleterious using the approach described above.

Permutation testing was used to evaluate the significance of whether multiple variants of either class were more or less frequent than expected. Because only genotype information was considered here, we assumed the independence of loci during the permutation process. Briefly, we randomly resampled genotypes for individuals in each variant and summarized the number of FA DA or DC variants found in each simulated individual. After 1000 replicates, we compared the observed distribution of DA/DC variants for each individual with the simulated distribution from permutation.

Conflict of Interest statement. None declared.

REFERENCES
