Genome-wide sequencing to identify the cause of hereditary cancer syndromes: with examples from familial pancreatic cancer

Nicholas J. Roberts¹ and Alison P. Klein²,³,⁺

¹Ludwig Center for Cancer Genetics, The Johns Hopkins University, Baltimore, MD 21231, USA
²Departments of Oncology and Pathology, The Johns Hopkins School of Medicine, Baltimore, MD 21231, USA
³Department of Epidemiology, Bloomberg School of Public Health, The Johns Hopkins University, Baltimore, MD 21231, USA

Abstract

Advances in our understanding of the human genome and next-generation technologies have facilitated the use of genome-wide sequencing to decipher the genetic basis of Mendelian disease and hereditary cancer syndromes. The application of genome-wide sequencing in hereditary cancer syndromes has had mixed success, in part, due to complex nature of the underlying genetic architecture. In this review we discuss the use of genome-wide sequencing in both Mendelian diseases and hereditary cancer syndromes, highlighting the potential and challenges of this approach using familial pancreatic cancer as an example.

1. Introduction

The publication of the draft human reference genome and advent of high-throughput sequencing technology have ushered in the genomics era. These new sequencing technologies have driven down the cost of genome sequencing to within a hair’s breadth of the $1000 dollar genome. Over a decade, per genome cost of sequencing has dropped precipitously, from approximately $70 million dollars in 2002, to nearly $7000 dollars in 2012 [1]. Similarly, the time to generate genomic sequence has also decreased, with over 10¹⁰ bases per instrument per day with the latest next generation sequencing technology. These advances resulted in the application of genome-wide sequencing to elucidate the underlying genetic basis of hereditary disease. Numerous examples of high-throughput, genome-wide approaches for disease gene identification are available for Mendelian diseases and these provide a starting framework for the analysis of complex diseases. Such gene discovery approaches share common elements, such as case selection and sample acquisition, genome-wide sequencing, data analysis to identify candidate genes with subsequent functional or population-based validation (Fig. 1). However, the application...
genomewide sequencing in the background of complex disease faces unique challenges due to underlying genetic heterogeneity of these diseases, in particular the identification of candidate genes using either filter-based or statistical based tests in the face of underlying genetic heterogeneity. In this review, we: 1) discuss the current use of genome-wide sequencing to identify the genetic basis of Mendelian diseases and hereditary cancer syndromes, 2) outline current knowledge about the genetic basis of familial pancreatic cancer, and 3) highlight the challenges of using genome-wide sequencing to identify pancreatic cancer susceptibility genes. We focus on familial pancreatic cancer because while genomic sequencing approaches have successfully identified familial pancreatic cancer genes, the challenges of using these approaches, specifically underlying genetic heterogeneity and its impact in study size requirements and candidate variant selection are faced by researchers using these approaches to study other cancer syndromes or other complex diseases

2. Genome-wide sequencing and hereditary disease

The first use of genome-wide sequencing to identify the genetic cause of a hereditary disease used Sanger sequencing to analyze the coding genes in the germline and tumor of a patient with familial pancreatic cancer [2]. Jones and colleagues employed a filter-based approach using the hypothesis that any susceptibility gene will be inactivating and heterozygous in the germline of the patient and contain a second mutation in the gene of the tumor, that is, it obeys the classical two-hit model of a tumor suppressor gene. Using this filter-based approach, 15,461 germline variants not seen in the human reference genome were narrowed to 3 in genes SERPINB12, RAGE and PALB2. SERPINB12 and RAGE were eliminated from further analysis as nonsense variants were common in healthy individuals. This left PALB2 as the putative susceptibility gene, highlighting the power of genome-wide sequencing to identify the genetic cause of a hereditary disease and the need of a good detective when approaching such a study.

Shortly after the discovery of PALB2 as a pancreatic cancer susceptibility gene, genomewide sequencing was successfully employed to identify the genetic basis of a Mendelian disorder. Ng and colleagues used next-generation technology to sequence the exome of 4 unrelated individuals with Freeman-Sheldon Syndrome, an autosomal dominant disorder of congenital arthrogryposis [3]. This study provided a powerful demonstration of the ability of next-generation technology to identify the genetic cause of a Mendelian disorder. Central to this proof-of-principle study were computational methods to find the single causative variant in the 20,000 rare variants per exome, the proverbial needle in the haystack. To do this, Ng and colleagues used a filter-based approach, eliminating common variants using control samples and the database of single nucleotide polymorphisms (dbSNP), and searching for genes where each affected individual harbored a rare variant. Their efforts identified one gene, MYH3, previously identified as the cause of Freeman-Sheldon Syndrome [4]. Subsequently, numerous Mendelian disorders where traditional linkage studies were unsuitable have been subjected to genome-wide sequencing approaches and the genetic basis of the phenotype defined [5; 6; 7; 8]. Such studies predominantly use whole-exome sequencing which limit search to exons and small segments of adjacent sequence, that is, coding and splice site variants. Fewer studies use whole-genome sequencing due to added cost and computational difficulties associated with a 20-fold increase in data per individual. However, studies utilizing whole-genome sequencing are not limited to coding variants and as such can find associations between copy number variation or non-coding variation and disease. A study by Jaeger and colleagues [9] used wholegenome sequencing in individuals with hereditary mixed polyposis syndrome (HMPS) to identify a duplication, 5’ to GREM1, that results in ectopic expression of GREM1 in colonic epithelium. As the authors note, further evidence for the causal nature of GREM1 duplications in HMPS patients is given by
juvenile polyposis syndrome (JPS). The majority of JPS patients harbor germline mutations in SMAD4 or BMPR1A [10], members of the bone morphogenetic protein (BMP) signaling pathway [11]. Interestingly, GREM1 is thought to negatively regulate BMP pathway signaling, providing a contemporaneous link between HMPS and JPS.

Hot on the heels of these successes were further large scale genome-wide sequencing studies to identify novel susceptibility genes in hereditary cancer syndromes. In 2012, Roberts and colleagues [12] reported on the whole-genome and whole-exome sequencing of 38 familial pancreatic cancer cases. Here they describe the aggregation of deleterious mutations in ATM in the germline of 6 pancreatic cancer cases representing 2 different kindreds. Similar to studies of Mendelian disorders described previously, Roberts and colleagues employ a filter-based approach using assumptions about the genetic architecture of susceptibility genes in familial pancreatic cancer to narrow down candidate genes for further study. In this study, germline variants present in the Human Gene Mutation Database (HGMD) [13] and associated with disease in people, were classified as functionally deleterious. This led to the identification of ATM as a candidate susceptibility gene and its validation in a larger sample set where 4 additional deleterious variants in 166 familial pancreatic cancer probands. Of note was the absence of deleterious ATM variants in a set of 190 controls which resulted in a statistically significant association of deleterious ATM variants in familial pancreatic cancer (Fisher’s exact test p-value = 0.046).

This was not the only successful application of genome-wide sequencing approaches to hereditary cancer syndromes. Park and colleagues [14] describe the use of whole-exome sequencing to identify XRCC2 as a breast cancer susceptibility gene. Again, a set of hereditary cases from 13 families, presumably enriched for susceptibility genes, were subjected to high-throughput sequencing, leading to the identification of a truncating and missense variant in the XRCC2 gene of 2 families. Validation in a case-control study of early onset breast cancer cases resulted in a statistically significant association of pathogenic coding variants in cases (6/1,308) vs. controls (0/1,120) (Fisher’s exact test p-value = 0.02). Interestingly, Park and colleagues used the computational algorithms: SIFT, Polyphen 2.1 and Align-GVGD, to assign functional significance to missense mutations. The contrast between the methods of assigning functional significance to missense variants by Roberts and colleagues and Park and colleagues are stark and occur due to the difficulty in assigning functional significance to the many rare and private germline variants discovered through genome-wide sequencing.

A study by Snape and colleagues [15] highlights perfectly the complexity of genetic risk determinants in hereditary cancer syndromes and the difficulty in identifying predisposition genes in hereditary breast cancers; even when an estimated 60 percent of familial risk for breast cancer is still to be explained [16]. Snape and colleagues conducted whole-exome sequencing in 50 individuals with familial breast cancer. Applying a filter-based approach and assumptions about the genetic basis of disease that were employed successfully in other genome-wide sequencing studies of hereditary cancer syndromes, Snape and colleagues failed to recognize susceptibility genes without a priori knowledge of breast cancer predisposition genetics. In their thorough study, an average of 26 truncating mutations per sample were identified and in 4 individuals the genetic basis of disease was assigned to truncating variants in three breast cancer susceptibility genes: BRCA2, CHEK2 and ATM. Validation of a subset of these truncating mutations with Sanger sequencing revealed a false-positive rate of 58.9 percent. However, once these false-positives were excluded from analysis, approximately 11 truncating variants per sample remained. Even in samples containing a truncating variant in a known susceptibility gene, other rare truncating variants occurred that may be linked to the disease phenotype e.g. regulators of apoptosis or transcriptional regulators, making interpretation difficult.
3. Lessons from hereditary pancreatic cancer

Genome-wide sequencing has proven useful with the identification of PALB2 and ATM as pancreatic cancer susceptibility genes. However, as described herein, the genetic basis of the overwhelming majority of familial cases remains undefined by convention linkage or candidate gene studies. Therefore, genome-wide sequencing provides an opportune approach to characterize elusive susceptibility genes essential for future genetic testing and targeted surveillance programs.

3.1. Evidence for a genetic basis to pancreatic cancer predisposition

Worldwide, more than 200,000 cases of pancreatic cancer are reported annually and the prognosis is particularly poor with a 5-year survival rate of less than 5 percent [17]. The disease is known to cluster in families, with 5–10 percent of cases in families with two first-degree relatives [18][19]. In 10–15 percent of these families, the genetic basis is attributed to known pancreatic cancer susceptibility, including: BRCA2, BRCA1, PALB2, STK11, CDKN2A, PRSS1, ATM, APC and mismatch repair genes (Table 1). However, the genetic basis in the remaining 85–90 percent of families with hereditary pancreatic cancer is unknown.

Co-occurrence of cases in a family may be due to underlying genetic basis, environmental risk factors or stochastic influences. Genetic contribution is thought to be the primary cause of aggregation [20] with evidence gleamed from family, twin, case-control and cohort studies. Initial case reports of familial clustering [21; 22; 23] were followed by observational epidemiological studies to assess the risk of pancreatic cancer in individuals with a family history of the disease. Fernandez and colleagues assessed 382 pancreatic cancers and 1408 controls and demonstrated excess risk of pancreatic cancer in individuals with a family history of the disease (relative risk 3.0) [24]. While the excess risk associated with a familial history of pancreatic cancer is beyond doubt, the magnitude of reported risk is variable between 2.3 and 32-fold increase [25]. Subsequently, the genetic contribution to pancreatic disease risk was highlighted and quantified by Lichtenstein and colleagues who surveyed a Swedish population-based twin registry, assessing disease concordance and discordance twins for a comprehensive list of cancers [26]. Estimates of the relative importance of heritable effects, the proportion of phenotypic liability attributable to genetic effects, was 0.36 (95 percent confidence interval: 0.00–0.53), suggesting a major role of genetic factors. However, the rarity of pancreatic cancer in the population prevented a statistically significant result. A population-based study of the Icelandic population, ascertained relative risk ratios for of different cancers for first, second, third, fourth and fifth-degree relatives of affected individuals [27]. For Pancreatic cancer, the risk ratio for first-degree relative of an affected was 2.3 (90 percent confidence interval: 1.8–3.0). This risk ratio was similar to other cancers assessed with a range of risk ratio for first degree relatives of 1.0 (90 percent confidence interval: 0.2–3.4) for myeloid leukemia and 5.0 (90 percent confidence interval: 2.8–9.5) for cancer of the lip. Interestingly, familial co-aggregation of pancreatic and ovarian cancer was significant (p-value < 0.05 percent) when considering first- to fifth-degree relatives. However, co-aggregation of pancreatic and breast cancers was not significant despite known susceptibility genes conferring risk to both breast and pancreatic cancer, for example, BRCA2. Similarly, a prospective study of pancreatic cancer families looked at the difference in observed and expected incidence of pancreatic cancers in families with at least one individual with pancreatic cancer. The ratio of observed to expected incidence of pancreatic cancers was 9.0 (95 percent confidence interval: 4.5–16.1) in familial pancreatic cancer kindreds versus sporadic pancreatic cancer kindreds with an observed-expected incidence ratio of 1.8 (95 percent confidence interval: 0.2–6.4) [18]. To quantify the magnitude of risk of pancreatic cancer in people with a family history of the disease, Wang and colleagues developed a Bayesian prediction model (PancPRO) has been
developed to estimate the probability of carrying a pancreatic cancer susceptibility gene and the absolute risk of pancreatic cancer in people with a family history of the disease [18]. PancPRPO was validated using data from 6,134 individuals, across 961 families of the National Familial Pancreatic Tumor Registry (NF PTR). PancPRO allows for the selection of a high risk subset of individuals for more intense disease surveillance.

3.2 Pancreatic cancer susceptibility genes

3.2.1 BRCA2, PALB2 and BRCA1—BRCA2, PALB2 and BRCA1 are important components of DNA repair pathways [28; 29]. Germline mutations of BRCA1 and BRCA2 are the underlying cause for hereditary breast and ovarian syndrome, an autosomal dominant condition with a lifetime risk of up to 80 percent for breast cancer [30]. PALB2 mutations have been associated with hereditary breast cancer [31]. In addition to breast and ovarian cancers, families with a germline BRCA2 mutation have a 3.5 fold (95 percent confidence interval: 1.9–6.6) risk of pancreatic cancer [32]. In familial pancreatic cancer families, 6–17 percent may harbor a germline BRCA2 mutation, making germline alterations in BRCA2 the most common identifiable cause of familial clustering of pancreatic cancer [33; 34]. As described previously, PALB2 was identified using genome-wide sequencing [2] and mutations have been reported in 1–3 percent of familial pancreatic cancer kindreds [35; 36] with carriers having an increased cancer risk of 5.9-fold (95 percent confidence interval: 2.4–14.6) [37]. The risk of pancreatic cancer in carriers of BRCA1 germline mutations is less well-established, while some studies suggest a 2.2-fold increased risk of pancreatic cancer in BRCA1 families [38], other studies fail to report an association [39; 40; 41]. Still, molecular typing of BRCA1, BRCA2 and PALB2 status in pancreatic cancers may be provide therapeutic benefit. BRCA1, BRCA2 and PALB2 deficient cancers are more susceptible to chemotherapeutic agents that disrupt DNA [42] and Poly(ADP-ribose) polymerase inhibitors [43; 44]. Recent clinical trials of PARP inhibitors in breast and ovarian cancer patients indicate that treatment was well tolerated, with durable objective responses seen in ~40 percent of cases [45; 46; 47]. Clinical trials are underway to examine the effects of treatment with PARP inhibitors in BRCA1, BRCA2 or PALB2 deficient pancreatic cancers and may provide new treatment options for this subgroup of patients.

3.2.2 STK11/LKB1—Germline mutations in STK11/LKB1 are associated with the autosomal dominant condition Peutz-Jeghers Syndrome (PJS) [48]. PJS patients are at increased risk of a variety of neoplasms, including: lung, breast, pancreatic, colon, gastric and ovarian [49]. Lifetime risk for any neoplasm in PJS patients is greater than 85 percent, with the relative risk of pancreatic cancer estimated to be increased 132-fold (95 percent confidence interval: 44–261) [49; 50].

3.2.3 CDKN2A—The gene CDKN2A is a tumor suppressor gene encoding p16^{INK4a} and p14^{ARF} [51]. Germline mutations in CDKN2A that disrupt p16^{INK4a} function result in aberrant cell cycle progression and are associated with familial atypical multiple mole melanoma (FAMMM) syndrome, an autosomal dominant condition accounting for approximately 40 percent of hereditary melanoma [52]. The link between FAMMM syndrome and pancreatic cancer was first identified by Lynch and colleagues [53]. The rate of pancreatic cancer in CDKN2A families is variable with up to 60 percent of kindreds being affected. Interestingly, this may be due to differential risk of pancreatic cancer associated with different CDKN2A mutations [51]. Overall, the relative risk of pancreatic cancer in carriers of CDKN2A mutations is thought to be 32-fold (95% confidence interval: 1.5–47.4) [54].

3.2.4 PRSS1 and SPINK1—Germline mutations of PRSS1 and SPINK1 are the cause of autosomal dominant and autosomal recessive hereditary pancreatitis respectively [55].
PRSS1 encodes cationic trypsinogen, a component of pancreatic juice and germline alterations, including: missense mutations, copy number alterations and microduplication, all perturb protein function through either increased trypsin stability or enhanced trypsinogen autoactivation [56]. Aberrant trypsin activity results in early-onset, severe, chronic pancreatitis, leading to permanent loss of endocrine and exocrine function and an increased risk of pancreatic cancer. Those affected by hereditary pancreatitis have up to an 80 fold increase risk of developing pancreatic cancer, with nearly 40 percent of individuals developing the disease by 70 years of age [57]. Similarly, SPINK1 encodes a trypsin inhibitor where biallelic germline mutations affecting expression or protein function removes a key checkpoint guarding against inappropriate trypsin activity, resulting in a chronic pancreatitis [56].

3.2.5 ATM—The ATM protein is a serine/threonine kinase integral to the repair of double strand breaks in DNA [58; 59]. Inheritance of biallelic deleterious variants in the ATM gene result in ataxia-telangiectasia [59], a disease that is characterized by progressive cerebellar ataxia, oculomotor apraxia, telangiectasias of the conjunctiva and skin, immunodeficiency, sensitivity to ionizing radiation and an increased rate of malignancies, in particular lymphoma and leukemia [58; 60; 61]. Population and family based studies point to an increased risk of breast and pancreatic cancers in carriers [12; 60; 61; 62]. Although, the magnitude of this effect is controversial [63; 64; 65; 66] and deleterious ATM variants are relatively common in the population, occurring in 0.5–1 percent of people [58; 60]. Further study is required to delineate the effect of environmental factors and modifier genes on disease risk in carriers of deleterious ATM variants.

3.2.6 Familial adenomatous polyposis and Lynch syndrome—Familial adenomatous polyposis (FAP) and Lynch syndrome (hereditary non-polyposis colon cancer – HNPCC), are associated with a moderately increased risk of pancreatic cancer. FAP, the archetypal colorectal cancer syndrome, is found in people harboring a germline, inactivating mutation in APC. Similarly, Lynch syndrome is an autosomal dominant condition of colorectal cancer predisposition caused by germline mutations in mismatch repair genes (MLH1, MSH2, MSH6 and PMS2) [67]. Besides an increased risk of colorectal cancer, patients with Lynch syndrome and FAP are more likely to develop certain extra-colonic malignancies, such as tumors of the: endometrium, pancreas, urinary tract, liver, thyroid and central nervous system [68; 69]. Recent studies have suggested the risk of pancreatic cancer in Lynch syndrome mutation carriers to be 3.7 percent (95 percent confidence interval: 1.6–5.9) up to age 70 years, represent a 8.6-fold increased risk (95 percent confidence interval: 4.7–15.7) [70]. This risk of pancreatic cancer associated with FAP is less clear while some studies fail to observe and increased risk of pancreatic cancers in FAP kindreds, up to a 4-fold increased risk have been observed in others [71].

3.3 Unexplained pancreatic cancer families

The vast majority 85–90 percent of familial pancreatic cancer kindreds are not explained by mutations in any of the above-mentioned genes. Given the rapid fatality of pancreatic cancer, DNA is often not available on multiple affected members of the same kindred. The lack of key DNA samples along with the underlying genetic heterogeneity, limits the success of linkage approaches to map susceptibility genes. Therefore, genomic sequencing approaches (Fig. 1) provide the best opportunity to discover novel pancreatic cancer susceptibility genes as evidenced by the recent identification of PALB2 and ATM using these approaches. However, genomic sequencing approaches provide a unique set of challenges.
4. Challenges of genome-wide studies of Hereditary Cancers

4.1 Multiple variants – nonsense and high-quality missense variants

The major challenge of using genomic sequencing to identify the causes of hereditary diseases is recognition of the causal mutation amidst the 1,000s of genetic variants identified, the needle in the haystack. On average, 35,000 and 6,000,000 variants are identified per exome or genome sequenced [12]. These will represent true variation, PCR or sequencing related errors, and bioinformatics artifacts. While majority of the changes may be excluded from further analysis as either common variants present in known SNP databases or observational artifacts below physical coverage thresholds, the number of rare and private variants is still far too large to be examined using functional approaches. Numerous statistical approaches have been developed to help identify the causal variant [72; 73], however, most of these approaches have not been applied in gene discovery studies. Furthermore, the power of these approaches will depend on the underlying biology, which is often unknown. The approaches proposed fall into one of two categories: filter-based approaches and statistical approaches commonly referred to as “burden tests”.

4.2 Filter-based approaches

To date, the sequencing studies that have successfully identified the causes of hereditary diseases, such as those outlined at the beginning of this review, primarily relied upon the filter-based approaches. As the name states, this approach uses a series of filters to eliminate benign variation resulting in a narrowed listed of potentially disease associated variants. Commonly used filters include variant type e.g. nonsense, missense and synonymous, allele frequency in publically available datasets e.g. HapMap, 1000 genomes [74; 75], and for missense variants, predicted function. Numerous functional prediction algorithms have been developed with the goal of discriminating between benign and functional genetic variation. These include SIFT, PolyPhen-2, GERP and others [76; 77; 78]. However, these algorithms have imperfect sensitivity and specificity [79]. Family data can also be used to filter-out variation by requiring variant-sharing among affected pedigree members and/or the absence of the variant in unaffected pedigree members.

4.3 Tumor sequencing

In addition to the above-mentioned approaches, knowledge of the underlying biology of the disease can also be used to help filter variants. When studying cancer syndromes, sequence analysis of the tumor itself can be especially powerful. Analysis of the tumor data not only can provide insight into the somatic mutations involved in tumor initiation and progression but also provide information on germ-line mutations. Data quality for tumor analysis provides its own set of challenges and is beyond the scope of this article, readers are directed to see Ding and Colleagues for a comprehensive review [80]. Given that tumors are very heterogeneous and represent a mixture of tumor and normal cells, identification of somatic mutations can be challenging. However, high-quality paired data from both the tumor and germline of a patient with hereditary cancer can be used to identify potential tumor suppressor genes by looking for genes in which there is a germline mutation coupled with either a somatic mutation of LOH in the cancer. While this approach may be ideal, at a minimum it doubles the cost and sequencing effort required. Furthermore, the amount of high quality tumor DNA needed for sequencing is often unavailable for most patients. If limited tumor DNA is available, tumor analysis can be limited to a small list of candidates. Tumors from patients carrying a variant in a given candidate can be analyzed for a second somatic mutation of LOH in the gene of interest. If paired data is not available, data from large-scale tumor sequencing such as those that have been conducted for breast, colon and pancreatic cancer [81; 82; 83; 84] can also be used as a filter. Variants that occur in genes that have been found to be somatically altered in other studies can be prioritized for follow-
up either in functional studies or in large-scale candidate genes studies. The major limitation to all of these filtering approaches is that the true causative mutation can easily be filtered-out and/or a very large number of potentially deleterious variants will remain after all filters have been used.

4.4 Burden tests

In contrast to the filtering based approaches, burden tests are a group of statistical tests aimed at identification of disease-associated variation. Traditional tests of association are not well suited to the analysis of sequencing data, as they typically require modest to large sample sizes. Because the frequency of any given variant in a gene can be quite low, these traditional association tests have little power to detect disease-causing mutations. To overcome this limitation, collapsing approaches or “burden-tests” have been developed. In brief, these methods combine variants detected in a given gene, pathway or region and then compare the frequency of these variants in cases and controls. A variety of burden test statistics have been developed. The Cohort Allelic Sums Test (CAST) groups all rare variants together [73], and the Combined Multivariate and Collapsing Method (CMC) groups rare variants and jointly examines them along with common variants [72]. These methods equally weight rare variants and thus assume all rare variants have a similar effect on the disease phenotype. Alternatively, variants can be weighted. While the best weighting strategy remains unclear and will likely depend on the underlying biology of the disease, common weighting strategies include the inverse of the allele frequency, functional information from prediction algorithms such as SIFT etc…, or estimates of genetic fitness.

4.5. Locus heterogeneity and incomplete penetrance of variants

The etiology of hereditary cancer syndromes can be quite complex with the majority of hereditary cancer syndromes exhibit locus heterogeneity, incomplete penetrance and phenocopies [85]. Locus heterogeneity occurs when mutations in more than one gene can lead to the same hereditary cancer phenotype. As locus heterogeneity increases the proportion of individuals sequenced who are likely to carry a disease causing mutation decreases. Because genome sequencing studies look for evidence of a mutation in a given gene across families, the presence of locus heterogeneity greatly limits the ability to identify disease associated genes. Similarly, not all individuals who inherit a mutation in hereditary cancer genes will go on to develop disease. While the probability of disease often increases with increasing age, even by age 80 a significant proportion of individuals may be disease free. For example, it is estimated that 29 and 54 percent of BRCA2 carriers develop breast cancer by age 50 and 70 respectively [86]. Because of incomplete penetrance, the interpretation of data from “unaffected” members of hereditary cancer families in genome sequencing studies must be carefully considered. Furthermore, many sequencing studies rely on publically available databases of “control” individuals. These controls represent either individual’s without disease, or in the case of 1000 genomes data [75] individuals with unknown cancer status. Therefore, it would not be unexpected to observe disease-causing variants in these control populations. Finally, sporadic cancers or phenocopies are not uncommon in hereditary pancreatic cancer families. These are cancers that develop in family members who did not inherit the disease predisposing mutations. Therefore, not all affected relatives will carry the disease causing mutations. Because the phenotype of the sporadic cancer mimics that of the hereditary cancer these phenocopies cannot be excluded a priori. The probability of a phenocopy depends on the frequency of the cancer in the general population.
5. Conclusion

Recent technological advancements allow for the broad application of genomic sequencing to discovery of the causes of hereditary diseases, including inherited cancer syndromes. While these methods are extremely powerful, as demonstrated by the identification of PALB2 and ATM as hereditary pancreatic cancer genes, identification of the single genetic change that leads to a hereditary cancer amidst the 1,000 of genetic changes identified though this approach can be elusive. For complex phenotypes, such as many hereditary cancers, locus heterogeneity, decreased penetrance and a high photocopy rate further complicate the interpretation of the sequence data. Despite these challenges, genome sequencing approaches will lead to the discovery of many cancer predisposition genes in near future.

References

1. Wetterstrand, KA. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. Available from: www.genome.gov/sequencingcosts


Figure 1. Schematic representation of cancer susceptibility gene discovery with genome-wide sequencing

Discovery relies on careful case identification and collection of matched tumor and normal samples. Genome-wide sequencing is conducted on a cohort of samples and bioinformatic analyses performed to identify individual variation. Candidate selection carried out with either filterbased approaches or burden tests before functional or population-based validation. Validated cancer susceptibility genes may be used for screening, risk assessment, prognostics and therapeutic targets.
Table 1

Known pancreatic cancer susceptibility genes and magnitude of their effect.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Fold increased risk of pancreatic cancer</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA2</td>
<td>3.5</td>
<td>[32]</td>
</tr>
<tr>
<td>BRCA1</td>
<td>2.2</td>
<td>[38]</td>
</tr>
<tr>
<td>PALB2</td>
<td>5.9</td>
<td>[37]</td>
</tr>
<tr>
<td>ATM</td>
<td>5</td>
<td>[12]</td>
</tr>
<tr>
<td>CDKN2A</td>
<td>32</td>
<td>[54]</td>
</tr>
<tr>
<td>STK11/LB1</td>
<td>132</td>
<td>[50]</td>
</tr>
<tr>
<td>PRSS1</td>
<td>80</td>
<td>[57]</td>
</tr>
<tr>
<td>MLH1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSH2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSH6</td>
<td>8.6</td>
<td>[70]</td>
</tr>
<tr>
<td>PMS2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APC</td>
<td>4.46</td>
<td>[71]</td>
</tr>
</tbody>
</table>