Title
Genetic Risk Scores and Missing Heritability in Ovarian Cancer

Permalink
https://escholarship.org/uc/item/4mv9m2rc

Journal
Genes, 14(3)

ISSN
2073-4425

Authors
Fatapour, Yasaman
Brody, James P

Publication Date
2023

DOI
10.3390/genes14030762

Copyright Information

Peer reviewed

Opinion

Genetic Risk Scores and Missing Heritability in Ovarian Cancer

Yasaman Fatapour and James P. Brody*[1]

[1]Department of Biomedical Engineering, University of California, Irvine.

* Correspondence: jpbrody@uci.edu.

Abstract: Ovarian cancers are curable by surgical resection when discovered early. Unfortunately, most ovarian cancers are diagnosed in the later stages. One strategy to identify early ovarian tumors is to screen women who have the highest risk. This opinion article summarizes the accuracy of different methods used to assess the risk of developing ovarian cancer, including family history, BRCA genetic tests, and polygenic risk scores. The accuracy of these is compared to the maximum theoretical accuracy, revealing a substantial gap. We suggest that this gap, or missing heritability, could be caused by epistatic interactions between genes. An alternative approach to computing genetic risk scores, using chromosomal-scale length variation should incorporate epistatic interactions. Future research in this area should focus on this and other alternative methods of characterizing genomes.

Keywords: copy number variation; ovarian cancer; machine learning; germline; UK Biobank; TCGA

Background.

Ovarian cancer is known as the silent killer. The symptoms of ovarian cancer in the initial stages are minimal and non-specific. Constipation, heartburn, fatigue, and bloating are early signs of ovarian cancer, but also associated with other common maladies. Because of these non-specific symptoms, ovarian cancer is often undiagnosed until the tumor has grown large, spread to nearby organs, and invaded the lymph system. At these later stages, treatment options are limited, and so is survival time. Ovarian tumors, like most solid tumors, can be surgically removed if found early. Removal of the tumor often leads to a complete cure[1]. However, most early detection strategies for ovarian cancer are ineffective for screening average risk women [2].

Current risk assessment tools for ovarian cancer do not work well enough. Specific genetic tests on BRCA1/BRCA2 status are available and work well for ovarian cancer, but only a small fraction (about 10%) of ovarian cancers are associated with those variants [3]. Otherwise, risk assessment is usually based on family history, but many people have limited knowledge of their family history and in any case a germline genetic test should work better than a perfect family history. Development of a genetic test to identify women at high-risk of ovarian cancer could lead to a reduction in the number of ovarian cancer deaths.

Genetic mutations are known to cause ovarian cancer, but the full extent is not known. The BRCA1/2 mutations account for a small percentage of ovarian cancer, but the others are called sporadic, with no known genetic cause. We suggest that these sporadic ovarian cancers are caused by as yet unknown genetic alterations in the germline. These genetic alterations might consist of epistatic effects, multiple combinations of mutations, unlike the simple mutations present in BRCA1/2 that cause ovarian cancer. This suggestion is based on several lines of evidence. First, analysis of somatic mutational data from tumors suggest that these tumors take decades to develop[32]. Second, age specific incidence data suggests that most cancers originate during development[33]. Third, the current lack of a detailed search for epistasis in germlines of ovarian cancer patients. Finally, an alternative hypothesis, that cancers originate from exposure to environmental mutagens that cause point mutations in people is not supported by evidence despite decades of studies [34].

Quantifying the accuracy of predictive tests.

Predictive tests often produce a numerical score that can be a continuous value, for instance from 1-100. From this score, one has to choose a cutoff value to make a prediction, which is a binary choice. Parameters like the sensitivity, specificity, positive predictive value, and negative predictive value are all a function of both the test and the choice of a cutoff value. The best way to characterize such a predictive test is with a Receiver Operating Characteristic curve [4,5]. This curve represents all cut off values, and one can read the sensitivity and specificity for the test for a given cutoff value.

The area under the curve of the receiver operating characteristic curve, or AUC, characterizes different predictive tests. The AUC, sometimes called a c-statistic, reduces the receiver operating characteristic curve to a single number, which is useful for comparing different tests. However, the complete ROC curve can show that two tests with similar AUC are not equivalent in some instances. Thus it is always best to examine the ROC curve for a test when judging its effectiveness.

The AUC can vary from 0.5, which is equivalent to random guessing, to 1.0, which indicates a perfect test that is always correct. The AUC is equivalent to the accuracy, when the two classes have equal numbers. The AUC is insensitive to class imbalance.

One example that illustrates how a predictive test with a low AUC can still be effective is the BRCA1 test for breast and ovarian cancer. This test works very well but only in a small subpopulation. Although the AUC is small, the test is quite valuable for that subpopulation.

Theoretical Maximum Accuracy of an Ovarian Cancer Genetic Risk Score

The highest possible AUC for predicting ovarian cancer in women is about 0.99 [6]. The discriminative accuracy of a genetic test depends on two factors, the heritability and prevalence of the trait. The Nordic Twin Study measured the heritability of ovarian cancer at about 40% [7]. Based on this heritability measurement and the prevalence of ovarian cancer, an ovarian cancer genetic test could have a maximum discrimination accuracy (AUC) in excess of 0.99. A substantial gap exists between the current best genetic risk tests and what should be possible.

Predicting Risk: Family History

Understanding a patient's family history is the first step in predicting whether a woman will develop ovarian cancer. Predictions based solely on family history have not been well characterized for ovarian cancer, but breast cancer predictive models have been well characterize, because it occurs ten times more frequently than ovarian cancer. For instance, one commonly used predictive model, the Gail model [8], has an **AUC of 0.58** (95% confidence interval [CI]=0.56 to 0.60) [9]. The Gail model incorporates several parameters including first degree relatives who were diagnosed with breast cancer but does not include any genetic information. Certain germline mutations in BRCA1 and BRCA2 are known to increase the risk of ovarian cancer.

The Tyrer-Cuzick model includes a more detailed picture of genetics including BRCA1/BRCA2 status and a hypothetical low-penetrance gene that is designed to encompass all other genetic factors [10]. The Tyrer-Cuzick model is an improvement over the Gail model and has an **AUC = 0.62**, with a 95% CI of (0.60 to 0.64) [11].

Several mutations in the BRCA1/BRCA2 genes are known to increase the risk of developing ovarian cancer. However, these mutations account for only about 10% of ovarian cancers in the general population[3,12]. Similarly, the fraction of breast cancers attributable to mutations in BRCA1 or BRCA2 is about 10%. Thus, the best AUC which could be expected for ovarian cancer predictive tests based on family history and supplemented with information on BRCA1/BRCA2 mutation status is probably similar to breast cancer, or about **AUC=0.60-0.65**[13–19].

The BRCA1/2 genetic tests are used to predict women at a high risk for breast and ovarian cancers. Some women whose BRCA test indicates a high risk of breast cancer choose to surgically remove their breasts to avoid breast cancer. Although less common, some women also choose a prophylactic oophorectomy--the surgical removal of the ovaries--to avoid ovarian cancers.

A positive BRCA1/2 test is highly predictive of breast/ovarian cancer, but a negative test is not very predictive of not having these cancers. In the US, only about 5-10% of breast and ovarian cancers are associated with mutations in BRCA1/2. A need exists to develop an effective genetic test for these other 90-95% of breast and ovarian cancers.

Predicting Risk: Polygenic Risk Scores

104 To fill this need, the most common approach is to use polygenic risk scores [13–19] . These
105 are linear combinations of single nucleotide polymorphisms (SNPs) found more often in
106 breast/ovarian cancer patients than in the general population. Models based on detailed germline
107 genetics should perform better than models based on family history alone, since family history is
108 often incomplete; limited to just a generation or two, and genetic factors present in relatives might
109 not be inherited.

110 The polygenic risk scores used today originate from Genome Wide Association Studies
111 (GWAS)[20–22]. These GWAS studies were designed to find genes that drive disease, not for
112 predictive tests. These polygenic risk scores are usually computed as a linear combination of the
113 "hits," each with a different weight, found in GWAS studies. Different algorithms use slightly
114 different criteria to decide on which "hits" to include and how to weigh them.

115 The current state of research knowledge on ovarian cancer genetic risk scores is best
116 represented by two recent papers. The first was published in *JNCI* in 2020 [23] and the second
117 was published in the *European Journal of Human Genetics* in 2022 [24].

118 The 2020 paper [23] evaluated polygenic risk scores for ovarian cancer, and seven other
119 common cancers, using the UK Biobank. In this dataset, they identified 358 women who had been
120 diagnosed with ovarian cancer. They constructed a polygenic risk score based upon 31 different
121 SNPs. Then, they evaluated the performance of this polygenic risk score to predict ovarian
122 cancer using the UK Biobank dataset. This test had a predictive accuracy of **AUC=0.568 (95%
123 CI 0.537 to 0.598)**.

124 The second paper, with over 150 authors, is a *tour-de-force* [24]. Compared to the first
125 paper, they increase the number of ovarian cancer subjects by nearly a factor of 100, using 23,564
126 cases. They thoroughly explored different combinations of SNPs and different algorithms for
127 combining these SNPs into a polygenic risk score. The second paper [24] describes the best
128 model found to be one based on measurements of 27,240 SNPs, almost 1000 times more than the
129 2020 paper [23]. After all that optimization, they found an **AUC of 0.588**. (They did not report a
130 95% confidence interval for the AUC).

131 Comparing the two papers, one can see that despite the extraordinary efforts of the second
132 paper, the AUC of the test was not significantly higher than the first paper (**AUC=0.588** vs **95%
133 CI 0.537 to 0.598**). From this comparison, we can conclude that most of the useful information
134 for predicting ovarian cancer has been extracted from SNP data using current algorithms. It seems
135 unlikely that the AUC can be significantly improved with different algorithms, a different set of
136 SNPs, or more patients in a dataset. This AUC is substantially lower than the theoretical
137 maximum; something is missing.

Missing heritability?

139 Many human diseases, including ovarian cancer, are known to be inherited. It was thought
140 that the advent of large scale genome wide association studies (GWAS) would reveal the
141 underlying genes that led to this inheritance for different disease[25,26]. However, GWAS results
142 have consistently shown that a substantial gap exists between the heritability that could be
143 attributed to known factors by GWAS and the heritability observed by studying inheritance in
144 families. The size of this gap varies by disease or trait, but it can be as large at a factor of ten [27].
145 The general missing heritability problem, and potential solutions, is well described by [26], in the
146 specific case of ovarian cancer, Flaum *et al* put it succinctly: "However, a significant proportion
147 of women who develop ovarian cancer with a strong family history of breast and/or ovarian
148 cancer still do not have a known variant to explain their increased risk, and there must be other
149 genetic factors at play that we do not yet understand."[12]

Beyond polygenic risk scores.

151 Epistatic interactions, nonadditive interactions between two or more genes, are one factor
152 usually cited as part of the missing heritability problem[26,28]. The methods used in GWAS
153 studies ignore non-linear interactions between genes, which are necessary to measure epistatic
154 interactions. Modern statistical techniques, or machine learning, allow one to consider non-linear
155 interactions between features, but these techniques inevitably require substantially more features
156 (SNPs) than samples (patients), which is not useful when a few thousand patient samples is
157 considered large, and genomes are characterized by millions of SNPs.

158 One approach to the problem is to construct a different representation of the genome as an
159 alternative to SNPs. A more compact representation that still accounts for the variability in
160 humans would allow the use of machine learning algorithms.

161 One example of this approach is to use measures of chromosome-scale length variation[29].
162 Chromosome-scale length variation can be computed from SNP array data. SNP arrays provide
163 calibrated intensity values for each SNP location. This intensity data is usually processed into

164 copy number variation data, which is represented by a multiplicity number (where two is the
165 normal multiplicity) and chromosome segment. Instead, one can take this intensity data and
166 compute an average multiplicity across an entire chromosome. By measuring this multiplicity
167 across an entire chromosome for many people, one finds a distribution in values (See Figure 1). A
168 person's germline genome, then, can be characterized by a series of twenty-three numbers where
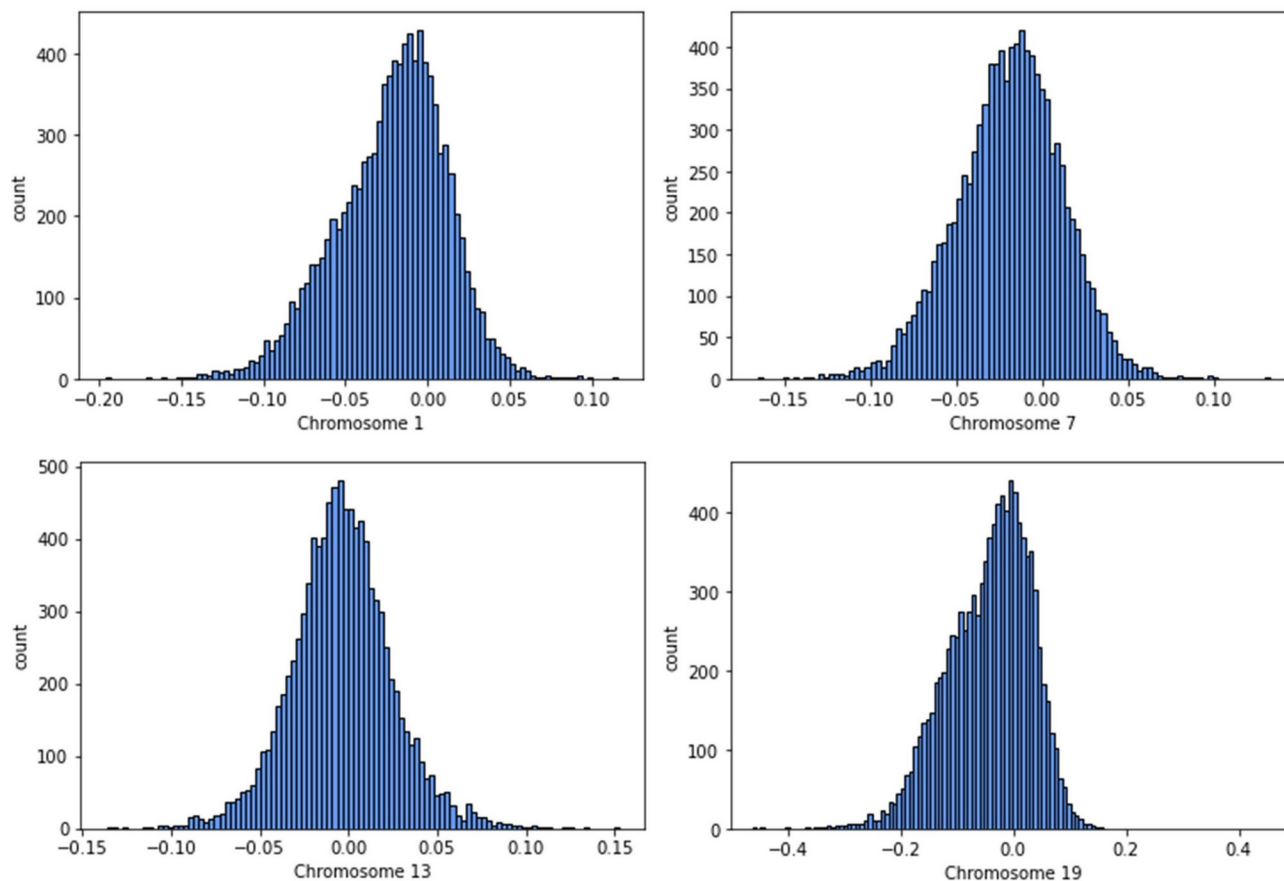169 each number represents the average multiplicity across each chromosome.

170

171 **Figure 1.** This figure shows a histogram of chromosome scale length variation measurements of
172 Chromosomes 1, 7, 13, and 19 for 10,000 people in the NIH All of Us dataset. "Chromosome length" is
173 measured by averaging calibrated intensity measurements taken from SNP arrays for many SNPs located on
174 each of the four chromosomes. These calibrated intensity measurements are representative of local copy
175 number. Chromosomes can have many deletions, insertions, and translocations that affect copy number.
176 The values measured in $\log_2$(Ratio Units) represent the overall length of the chromosome, where a value of
177 zero indicates the nominal average chromosome length. By measuring this parameter for all chromosomes,
178 one can characterize each person's germline genetic makeup with these 23 numbers.

179 This representation of a person's genome, twenty-three numbers, has some advantages over
180 the conventional SNP representation of a genome. It is more compact, but still sufficiently
181 complex to capture the enormity of the human population. The compactness allows one to use
182 modern machine learning techniques. It is extensible; you can split the chromosomes into
183 arbitrarily small sections.
184 Using a data set acquired as part of the Cancer Genome Atlas (TCGA) project, we evaluated
185 a genetic risk score computed from chromosomal scale length variation derived from TCGA
186 normal blood samples. In this data set, it had an AUC of 0.88 (95% CI of 0.86-0.91)[29]. Women
187 with the highest 20% had 160 times the risk of developing ovarian cancer as compared to the
188 lowest 20%. Although these numbers showed extraordinary discrimination, it is unclear whether
189 these results are generalizable to the general population. The TCGA data set only contains people
190 who had been diagnosed with cancer, so this work really distinguished one form of cancer from
191 other forms of cancer. It is also possible that the TCGA has subtle batch effects, leading to
192 falsely high discrimination[30,31].

193 **Conclusion.**

194 Ovarian cancer is completely curable in the early stages. While convincing data does not yet
195 exist, we believe that the propensity to develop ovarian cancer appears to be transmitted through
196 the genome, primarily through epistatic interactions. Thus, our opinion is that identification of
197 signatures in the germline genome that indicate future diagnosis of ovarian cancer should be a
198 primary and important target of research. We describe one early effort to use chromosome scale
199 length variation measurements to quantify insertions and deletions that might hold promise for
200 predicting risk of developing ovarian cancer.

References

202 1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72.
203 2. Grossman DC, Curry SJ, Owens DK, Barry MJ, Davidson KW, Doubeni CA, et al. Screening for ovarian cancer US preventive services
204 task force recommendation statement. JAMA - Journal of the American Medical Association. 2018.
205 3. Ramus SJ, Gayther SA. The Contribution of BRCA1 and BRCA2 to Ovarian Cancer. Mol Oncol. 2009.
206 4. Moons KGM, de Groot JAH, Linnet K, Reitsma JBR, Bossuyt PMM. Quantifying the added value of a diagnostic test or marker. Clin
207 Chem. 2012.
208 5. Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the
209 performance of diagnostic tests. Clin Biochem Rev. The Australian Association of Clinical Biochemists; 2008;29 Suppl 1: S83-7.
210 6. Janssens ACJW, Aulchenko YS, Elefante S, Borsboom GJJM, Steyerberg EW, van Duijn CM. Predictive testing for complex diseases
211 using multiple genes: Fact or fiction? Genetics in Medicine. Nature Publishing Group; 2006; 8:395–400.
212 7. Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, et al. Familial Risk and Heritability of Cancer Among Twins in
213 Nordic Countries. JAMA. American Medical Association; 2016; 315:68
214 8. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting Individualized Probabilities of Developing Breast
215 Cancer for White Females Who Are Being Examined Annually. JNCI: Journal of the National Cancer Institute. Oxford Academic; 1989;
216 81:1879–86.
217 9. Chlebowski RT, Anderson GL, Lane DS, Aragaki AK, Rohan T, Yasmeen S, et al. Predicting risk of breast cancer in postmenopausal
218 women by hormone receptor status. J Natl Cancer Inst. 2007;99.
219 10. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med. John Wiley &
220 Sons, Ltd; 2004;23:1111–30.
221 11. McCarthy AM, Guan Z, Welch M, Griffin ME, Sippo DA, Deng Z, et al. Performance of Breast Cancer Risk-Assessment Models in a
222 Large Mammography Cohort. J Natl Cancer Inst. 2020;112.
223 12. Flaum N, Crosbie EJ, Edmondson RJ, Smith MJ, Evans DG. Epithelial ovarian cancer risk: A review of the current genetic landscape. Clin
224 Genet. Blackwell Publishing Ltd; 2020; 97:54–63.
225 13. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018; 19:581–90.
226 14. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. BioMed Central; 2020 12:44.
227 15. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Hum Mol Genet. 2019.
228 16. Hughes E, Tshiaba P, Gallagher S, Wagner S, Judkins T, Roa B, et al. Development and Validation of a Clinical Polygenic Risk Score to
229 Predict Breast Cancer Risk. JCO Precis Oncol. American Society of Clinical Oncology; 2020;585–92.
230 17. Khera A v., Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify
231 individuals with risk equivalent to monogenic mutations. Nat Genet. Nature Publishing Group; 50:1219–24.
232 18. Sugrue LP, Desikan RS. What Are Polygenic Scores and Why Are They Important? JAMA. American Medical Association; 2019
233 321:1820.
234 19. Li R, Zhang X, Li B, Feng Q, Kottyan L, Luo Y, et al. Polygenic risk vectors (PRV) improve genetic risk stratification for cardio-
235 metabolic diseases. medRxiv. 2022.
236 20. Goode EL, Chenevix-Trench G, Song H, Ramus SJ, Notaridou M, Lawrenson K, et al. A genome-wide association study identifies
237 susceptibility loci for ovarian cancer at 2q31 and 8q24. Nat Genet. 2010.
238 21. Evans DM, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual
239 prediction of complex disease risk. Hum Mol Genet. 2009
240 22. Reid BM, Permuth JB, Chen YA, Fridley BL, Iversen ES, Chen Z, et al. Genome-wide Analysis of Common Copy Number Variation and
241 Epithelial Ovarian Cancer Risk. Cancer Epidemiol Biomarkers Prev. NIH Public Access; 2019
242 23. Jia G, Lu Y, Wen W, Long J, Liu Y, Tao R, et al. Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for
243 Eight Common Cancers. JNCI Cancer Spectr. Oxford Academic; 2020
244 24. Dareng EO, Tyrer JP, Barnes DR, Jones MR, Yang X, Aben KKH, et al. Polygenic risk modeling for prediction of epithelial ovarian
245 cancer risk. European Journal of Human Genetics 2021 30:3. Nature Publishing Group; 30:349–62
246 25. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases.
247 Nature. 2009.
248 26. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of
249 complex disease. Nat Rev Genet. 2010.
250 27. Génin E. Missing heritability of complex diseases: case solved? Hum Genet. 2020.
251 28. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl
252 Acad Sci U S A. National Academy of Sciences; 2012; 109:1193–8.
253 29. Toh C, Brody JP. Genetic risk score for ovarian cancer based on chromosomal-scale length variation. BioData Min. BioMed Central Ltd;
254 2021;14.
255 30. Jain S, Mazaheri B, Raviv N, Bruck J. Short Tandem Repeats Information in TCGA is Statistically Biased by Amplification. bioRxiv.
256 2019.
257 31. Jain S, Mazaheri B, Raviv N, Bruck J. Glioblastoma signature in the DNA of blood-derived cells. PLoS One. 2021;16.
258 32. Beerenwinkel N, Antal T, Dingli D, Traulsen A, Kinzler KW, Velculescu VE, Vogelstein B, Nowak MA. Genetic progression and the
259 waiting time to cancer. PLoS Comput Biol. 2007 Nov;3(11):e225.

33. Brody, J. P. The Age Specific Incidence Anomaly Suggests that Cancers Originate During Development. Biophys. Rev. Lett. 2014 9(02):193-204

34. Thilly, W. G. Have environmental mutagens caused oncomutations in people? Nat Genetics 2003 34(3) 255-259.