Inherited factors are thought to be responsible for a substantial fraction of many different forms of cancer. However, individual cancer risk cannot currently be well quantified by analyzing germ line DNA. Most analyses of germ line DNA focus on the additive effects of single nucleotide polymorphisms (SNPs). Here we show that chromosomal-scale length variation of germ line DNA can be used to predict whether a person will develop cancer. In two independent datasets, the Cancer Genome Atlas (TCGA) project and the UK Biobank, we could classify whether or not a patient had a certain cancer based solely on chromosomal scale length variation. We found a model that could predict ovarian cancer in women with an area under the receiver operator curve (AUC) of 0.89. In the UK Biobank data, we could predict breast cancer in women with an AUC of 0.83.

We found a model that could predict ovarian cancer in women with an area under the receiver operator curve, known as the AUC, for ovarian cancer in women (0.89), glioblastoma multiforme (0.88), breast cancer in women (0.75), and colon adenocarcinoma (0.79), with a 95% confidence interval width of less than 0.01 in each case. This method could predict 10% of glioblastoma multiforme cancers and 10,000 false positive diagnoses in the TCGA dataset. In the UK Biobank dataset, we could predict breast cancer in women with an AUC of 0.83.

Here we show that chromosomal-scale length variation of germ line DNA can be used to predict whether or not a person will develop cancer based solely on chromosomal scale length variation. In two independent datasets, the Cancer Genome Atlas (TCGA) project and the UK Biobank, we could classify whether or not a person had a certain cancer based solely on chromosomal scale length variation. We found a model that could predict ovarian cancer in women with an area under the receiver operator curve, known as the AUC, for ovarian cancer in women (0.89), glioblastoma multiforme (0.88), breast cancer in women (0.75), and colon adenocarcinoma (0.79), with a 95% confidence interval width of less than 0.01 in each case. This method could predict 10% of glioblastoma multiforme cancers and 10,000 false positive diagnoses in the TCGA dataset. In the UK Biobank dataset, we could predict breast cancer in women with an AUC of 0.83.

Inherited factors are thought to be responsible for a substantial fraction of many different forms of cancer. However, individual cancer risk cannot currently be well quantified by analyzing germ line DNA. Most analyses of germ line DNA focus on the additive effects of single nucleotide polymorphisms (SNPs). Here we show that chromosomal-scale length variation of germ line DNA can be used to predict whether a person will develop cancer. In two independent datasets, the Cancer Genome Atlas (TCGA) project and the UK Biobank, we could classify whether or not a patient had a certain cancer based solely on chromosomal scale length variation. We found a model that could predict ovarian cancer in women with an area under the receiver operator curve (AUC) of 0.89. In the UK Biobank data, we could predict breast cancer in women with an AUC of 0.83. This method could be used to develop genetic risk scores for other conditions known to have a substantial genetic component and complements genetic risk scores derived from SNPs.

We tested this objective using data compiled by The Cancer Genome Atlas (TCGA) project and the UK Biobank. TCGA data consisted of information on DNA chromosome scale length variation extracted from the peripheral blood of 6,821 different patients, each of whom had one of 32 different types of cancer. The UK Biobank data consisted of about 1500 women diagnosed with breast cancer (cancer cases) and about 5000 women who have no record of any type of cancer (controls).

We characterized structural variation by a quantitative measure of chromosomal-scale length variation. Chromosomal-scale length variation is computed from array data in the TCGA data. It is a product of the copy number variation pipeline. In the UK Biobank data, we computed it as the average value of the CNV log2r data across the chromosome. We can characterize a person’s germ line DNA by a series of 22 numbers, each representing the log base 2 ratio of the chromosome’s length compared to the average chromosome length. Using these 22 numbers for each person, we set up a machine learning classification problem to differentiate those people diagnosed with a particular form of cancer from those who have not been diagnosed with that cancer. Similarly, we can break each chromosome into four equal parts and use 88 numbers for each person. We used the h2o libraries in R to test how well different machine-learning algorithms can classify whether a person had a specific cancer or not, based solely on these numbers derived from germ line DNA.

We tested this objective using data compiled by The Cancer Genome Atlas (TCGA) project and the UK Biobank. TCGA data consisted of information on DNA chromosome scale length variation extracted from the peripheral blood of 6,821 different patients, each of whom had one of 32 different types of cancer. The UK Biobank data consisted of about 1500 women diagnosed with breast cancer (cancer cases) and about 5000 women who have no record of any type of cancer (controls).

We characterized structural variation by a quantitative measure of chromosomal-scale length variation. Chromosomal-scale length variation is computed from array data in the TCGA data. It is a product of the copy number variation pipeline. In the UK Biobank data, we computed it as the average value of the CNV log2r data across the chromosome. We can characterize a person’s germ line DNA by a series of 22 numbers, each representing the log base 2 ratio of the chromosome’s length compared to the average chromosome length. Using these 22 numbers for each person, we set up a machine learning classification problem to differentiate those people diagnosed with a particular form of cancer from those who have not been diagnosed with that cancer. Similarly, we can break each chromosome into four equal parts and use 88 numbers for each person. We used the h2o libraries in R to test how well different machine-learning algorithms can classify whether a person had a specific cancer or not, based solely on these numbers derived from germ line DNA.

In two independent datasets, the Cancer Genome Atlas (TCGA) project and the UK Biobank, we could classify whether or not a person had a certain cancer based solely on chromosomal scale length variation. We found all 32 different types of cancer in the TCGA dataset tested could be predicted better than chance using structural variation data. Specifically, in the TCGA dataset we measured the area under the receiver operator curve, known as the AUC, for ovarian cancer in women (0.89), glioblastoma multiforme (0.88), breast cancer in women (0.75), and colon adenocarcinoma (0.79), with a 95% confidence interval width of less than 0.01 in each case. This method could predict 10% of glioblastoma multiforme cancers and 10,000 false positive diagnoses in the TCGA dataset. In the UK Biobank dataset, we could predict breast cancer in women with an AUC of 0.83.

Here we show that chromosomal-scale length variation of germ line DNA can be used to predict whether a person will develop cancer. In two independent datasets, the Cancer Genome Atlas (TCGA) project and the UK Biobank, we could classify whether or not a patient had a certain cancer based solely on chromosomal scale length variation. We found all 32 different types of cancer in the TCGA dataset tested could be predicted better than chance using structural variation data. Specifically, in the TCGA dataset we measured the area under the receiver operator curve, known as the AUC, for ovarian cancer in women (0.89), glioblastoma multiforme (0.88), breast cancer in women (0.75), and colon adenocarcinoma (0.79), with a 95% confidence interval width of less than 0.01 in each case. This method could predict 10% of glioblastoma multiforme cancers and 10,000 false positive diagnoses in the TCGA dataset. In the UK Biobank dataset, we could predict breast cancer in women with an AUC of 0.83.

In two independent datasets, the Cancer Genome Atlas (TCGA) project and the UK Biobank, we could classify whether or not a person had a certain cancer based solely on chromosomal scale length variation. We found all 32 different types of cancer in the TCGA dataset tested could be predicted better than chance using structural variation data. Specifically, in the TCGA dataset we measured the area under the receiver operator curve, known as the AUC, for ovarian cancer in women (0.89), glioblastoma multiforme (0.88), breast cancer in women (0.75), and colon adenocarcinoma (0.79), with a 95% confidence interval width of less than 0.01 in each case. This method could predict 10% of glioblastoma multiforme cancers and 10,000 false positive diagnoses in the TCGA dataset. In the UK Biobank dataset, we could predict breast cancer in women with an AUC of 0.83.

In two independent datasets, the Cancer Genome Atlas (TCGA) project and the UK Biobank, we could classify whether or not a person had a certain cancer based solely on chromosomal scale length variation. We found all 32 different types of cancer in the TCGA dataset tested could be predicted better than chance using structural variation data. Specifically, in the TCGA dataset we measured the area under the receiver operator curve, known as the AUC, for ovarian cancer in women (0.89), glioblastoma multiforme (0.88), breast cancer in women (0.75), and colon adenocarcinoma (0.79), with a 95% confidence interval width of less than 0.01 in each case. This method could predict 10% of glioblastoma multiforme cancers and 10,000 false positive diagnoses in the TCGA dataset. In the UK Biobank dataset, we could predict breast cancer in women with an AUC of 0.83.

In two independent datasets, the Cancer Genome Atlas (TCGA) project and the UK Biobank, we could classify whether or not a person had a certain cancer based solely on chromosomal scale length variation. We found all 32 different types of cancer in the TCGA dataset tested could be predicted better than chance using structural variation data. Specifically, in the TCGA dataset we measured the area under the receiver operator curve, known as the AUC, for ovarian cancer in women (0.89), glioblastoma multiforme (0.88), breast cancer in women (0.75), and colon adenocarcinoma (0.79), with a 95% confidence interval width of less than 0.01 in each case. This method could predict 10% of glioblastoma multiforme cancers and 10,000 false positive diagnoses in the TCGA dataset. In the UK Biobank dataset, we could predict breast cancer in women with an AUC of 0.83.

In two independent datasets, the Cancer Genome Atlas (TCGA) project and the UK Biobank, we could classify whether or not a person had a certain cancer based solely on chromosomal scale length variation. We found all 32 different types of cancer in the TCGA dataset tested could be predicted better than chance using structural variation data. Specifically, in the TCGA dataset we measured the area under the receiver operator curve, known as the AUC, for ovarian cancer in women (0.89), glioblastoma multiforme (0.88), breast cancer in women (0.75), and colon adenocarcinoma (0.79), with a 95% confidence interval width of less than 0.01 in each case. This method could predict 10% of glioblastoma multiforme cancers and 10,000 false positive diagnoses in the TCGA dataset. In the UK Biobank dataset, we could predict breast cancer in women with an AUC of 0.83.