bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Free-living human cells reconfigure their chromosomes in the evolution back to uni-cellularity

Jin Xu[1,2,6], Xinxin Peng[3,6,] Yuxin Chen[2,6], Yuezheng Zhang[3], Qin Ma[3,4], Liang Liang[3,4], Ava C. Carter[1], Xuemei Lu[3*] and Chung-I Wu[2,5*]

Author affiliation:
[1]Center for Personal Dynamic Regulomes and Program in Epithelial Biology, Stanford University School of Medicine, Stanford, California 94305, USA
[2]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou 510275, China.
[3]Laboratory of Genome Variations and Precision Biomedicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, P.R. China.
[4]University of Chinese Academy of Sciences, Beijing 100049, China
[5]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA
[6] These authors contributed equally to this study.
[*]Corresponding authors: E-mail: ciwu@uchicago.edu or luxm@big.ac.cn
Lead Contact: ciwu@uchicago.edu

Keywords: sex chromosome evolution; dosage compensation; multi-cellularity; uni-cellularity; cancer cell evolution

Abstract

Cells of multi-cellular organisms evolve toward uni-cellularity in the form of cancer and, if humans intervene, continue to evolve in cell culture. During this process, gene dosage relationships may evolve in novel ways to cope with the new environment and may regress back to the ancestral uni-cellular state. In this context, the evolution of sex chromosomes vis-a-vis autosomes is of particular interest. Here, we report the chromosomal evolution in ~600 cancer cell lines. Many of them jettisoned either Y or the inactive X; thus, free-living male and female cells converge by becoming "de-sexualized". Surprisingly, the active X often doubled, accompanied by the addition of one haploid complement of autosomes, leading to an X:A ratio of 2:3 from the extant ratio of 1:2. Theoretical modeling of the frequency distribution of X:A karyotypes suggests that the 2:3 ratio confers a higher fitness and may reflect aspects of sex chromosome evolution.

Introduction

Genomes of multi-cellular organisms evolve to ensure the survival and reproduction of the whole organisms. With human interventions akin to domestication, hundreds of cell lines survive as free-living cells that are not organized into tissues, organs or individuals[1]. Evolution in such a qausi-unicellular state may be very different from the evolution as multi-cellular entities. Most cell lines are cancerous in origin but a few are derived from normal tissues[2]. Regardless of their origin, they have all evolved characteristics for survival in the unicellular state that is distinct from their natural environments. Cell lines derived from cancer tissues are usually karyotypically less stable than normal cell lines. While this instability may impose a cost, it also permits cancer cell lines to evolve new karyotypes, including polyploidy, more readily than normal cell lines could.

Tumorigenesis has been increasingly viewed as a process of evolution, rather than merely pathological conditions[3][4]. This "ultra-microevolutionary process" is subjected to similar rules including mutation, genetic drift, migration and selection that govern organismal evolution[5]. While this process usually ends when the organism dies, cell lines in the cultured state will continue

bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to evolve. Much like the diversity unleashed by domestication, cultured cell lines, which can be considered "domesticated", may be informative about the evolutionary potentials at the cellular level.

In this quasi-unicellular state, gene dosage has been observed to change extensively as polyploidy, aneuploidy (full or partial) and various copy number variations (CNVs) are common in cancer cell lines[6]. Since these cells lines are derived from somatic tissues of man or woman (referred to as male and female cells, for simplicity), they should be different in their sex chromosomes in relation to the autosomes (A's). Nevertheless, the possibility of separate evolutionary paths has not been raised before. Somatic cells have an inactive X chromosome in females and a Y chromosome in males[7]. Since cell lines presumably do not need sexual characters, we ask how the X:A relationship might have evolved in both male and female cells. More generally, we ask whether the evolution in this relationship may shed light on the emergence of mammalian sex chromosomes and the subsequent evolution.

In this study, we analyze 620 cancer cell lines that have been genotyped using SNP arrays[8]. Among them, 279 are derived from female tissues and 341 from male tissues. We observed the elimination of the Y and the inactive X chromosome, followed by the evolution toward a new equilibrium with 2 active X chromosomes and 3 sets of autosomes (2X:3A). We discuss the implication of these findings for the evolution of sex chromosome, the transition between uni- and multi-cellularity and cancers biology.

Results

Convergent sex chromosomes evolution between sexes

A most common form of genomic changes in cell lines is the loss of heterozygosity (LOH) when one of the two homologous chromosomes is eliminated[6]. We therefore examine single nucleotide polymorphisms (SNPs) across the 620 cell lines for occurrences of LOH on each autosome and the X chromosome. Male and female cell lines are separately analyzed.

Figure 1A shows the LOH frequency for each autosome (black dots) and the red dot represents the sex chromosomes (X in female and Y in male). For autosomes, the percentages of LOH are remarkably similar between sexes, with a correlation coefficient of 0.94 among 620 cell lines. There is a slight tendency for the smaller autosomes to have higher LOH rate (R=~-0.4, p=~0.046, Figure S1). The median percentage of LOH is about 13% for autosomes. However, the losses of X (36% in females) and Y (40% in males) stand out. Given its rank as the 7th largest chromosome, the X is not expected to be lost in more than 15% of cell lines. Since the X expression is not lost, we infer that it's the inactive X(or Xi) that is eliminated.

Female lines lose the inactive X (Xi) and male lines lose the Y chromosome at a higher rate than other chromosomes. The two sexes may thus be expected to converge toward having a single sex chromosome. Furthermore, given that spontaneous LOH is not infrequent and the loss cannot be regained, long term cultures might evolve to complete LOH for sex chromosomes as well as autosomes. The genome-wide low rate of LOH suggests selection holding back such changes. The strong correlation between sexes further reflects a balance between the production and elimination of LOH's, likely involved natural selection.

A most unexpected finding is that, accompanying the loss of the Y or Xi, an extra X chromosome is often gained. **Figure 1B** shows approximately equal numbers of male cell lines with one or two X chromosomes (partial X aneuploidy not counted). This extra X is active because the inactivating

bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

XIST lncRNA is silenced in male cell lines (**Figure 1C**), consistent with previous findings[9]. XIST does not become activated in free-living cells that do not already express this. The expression of X-linked genes is higher in those male lines with two X's than in those with one X and the up-regulation occurs along the length of the X chromosome (**Figure 1D**).

The pattern is more complex in female lines which, in their original state, contain an Xa and an Xi, the latter expressing XIST[10][11][12]. We focus on female lines that experienced LOH of the X, which should be genetically equivalent to male lines that have lost the Y. These female lines indeed evolve in a manner identical with the male lines. First, female lines that have gained an X are almost as frequent as lines with one X, much like the male lines with one vs. two Xa's (**Figure 1E**). Second, female lines with an additional X do not express Xist and all X's can thus be presumed active (**Figure 1F**)[13]. As in male lines, the X does not switch its state after chromosome duplication.

Cancer cell lines usually have high rate of aneuploidy and could be heterogeneous within the line, thus making its status difficult to assess. To assess the level of within-line heterogeneity, we chose two representative cell lines to count the X chromosomes in individual cells using fluorescent in situ hybridization (FISH). The two lines are A549 (a male cell line from adenocarcinomic alveolar basal epithelium) and HeLa (a female cervical cancer cell line). Neither line expresses XIST (Table S2), suggesting that all X chromosomes are active. **Figure 2A-B** shows results from individual A549 and HeLa cells with two and three X's. **Figure 2C-D** shows the X karyotype distributions. While there is a modest degree of heterogeneity within each line, almost all cells have two or more active X chromosomes. While labor intensity of assays and cell availability limited our sample size, we nevertheless can conclude that within-cell line heterogeneity does not seem to undermine our conclusions.

Evolution toward a new X:A expression ratio ($E_{X/A}$)

With an extra copy of active X, the "expression phenotype" is expected to change. The ratio of the median gene expression on the X to that on the autosomes($E_{X/A}$) is of particular interest. $E_{X/A}$ has been reported to be around 0.5~0.8 for normal mammalian tissues[14–16]. We assayed $E_{X/A}$ by separating lines derived from cancerous and normal tissues. **Figure 3A** shows that $E_{X/A}$ distributions center on ~0.84 in normal cell lines and on 1 in cancerous cell lines. Given the controversy in the assay of $E_{X/A}$, we also varied the threshold for counting expressed transcripts (see Materials and Methods). By varying the threshold (**Figure 3B**), $E_{X/A}$ ranges from 0.78 to 1.05 in normal cell lines but is consistently higher by approximately 15% in cancer cell lines. The same pattern is seen in the RNA-seq data (Figure S2).

The concerted evolution of autosomes as a set

While sex chromosomes evolve, autosomes should also evolve. Since the generation of aneuploidy may happen independently for each autosome, a key question is whether selection operates on the autosomes as a set. Does natural selection favor cells that have full sets of autosomes?

Figure 4A shows the distribution of chromosome number across the 620 cell lines we studied. Apparently, cancerous cell lines acquire autosomes during evolution. The distribution of ploidy (n=22) number shows peaks at 2 and 3, indicates many cell lines appear to be in transition between full diploidy and triploidy of 44 and 66 autosomes. Similarly, the majority of sublines of HeLa cells we examined have 55-75 chromosomes centering about the triploid count of 69 (Figure S4A). Indeed, autosomes appear to exist as a full complement with n=22. Although autosomes may evolve as a set, cells most likely add one autosome at a time. It is hence desirable to track each chromosome

bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

individually. Single cells were individually isolated from a HeLa cell line and subsequently grown to a sub-line of 10^6 cells. We subjected 6 such sub-lines to whole genome sequencing such that each chromosome can be tracked individually. Smaller chromosomes are indeed more erratic in their numbers in cell lines. Only the largest 14 chromosomes (13 autosomes and X), which together account for ~75% of the genome, are used to test the convergence of autosomes. The cutoff is based on the observation that chromosome 13 is the largest autosome yielding viable trisomic new-borns[17–19]. We reason that, if whole organisms can survive trisomy, the fitness consequence of the particular aneuploidy would probably be very small at the cellular level.

In all 6 lines, each of the 13 autosomes has 2 – 4 copies, ranging from an average of 2.62 to 3.23 (Table S1). If each autosome behaves independently, the number of autosomes that increase by x copies (x = 0, 1, 2 etc.) should follow a Poisson distribution with a mean of λ. Two different lines, with λ = 10/13 and λ = 16/13, are shown in **Figure 4B and C**. In the former, all cells have x = 0 or x=1 and, in the latter, all cells have x =1 or x=2 (Table S1). The data suggest that each autosome increases by one copy and only after all of the 13 autosomes have gained an extra copy do further increases continue. Figure S4B shows the composite distribution of the five lines with λ < 1. The pattern, like that of **Figure 4B**, is statistically significant (P = 0.0021 by the χ^2 test) with an excess at x =1. These results suggest that the larger autosomes evolve cohesively as a set. With autosomes evolving as a cohesive unit, X:A can be represented by whole numbers of 1:2, 2:3 etc.

Evolution of the C(Xa:A) ratio underlying $E_{X/A}$

We now summarize the evolution of cell lines by their C(Xa:A) genotypes. C(Xa:A) is the number of active X chromosomes and the ploidy number of autosomes (in multiples of 22) and is equal to C(1,2) in normal cells. For the purpose of counting on active Xa's, data from most male lines are usable. For female lines, only data from the LOH lines of the X can be used. Between the two sexes, C(Xa:A) distributions are very similar and the combined distribution is used in the analysis (Figure S4C).

Shown in **Figure 4D**, most lines have the C(1:2) or C(2:3) genotype which together account for 2/3 of the lines. Given that C(1:2) is the starting genotype, its common occurrence at 37.4% is not surprising. The high frequency of C(2:3), however, is unexpected. To reach C(2:3) from the starting point of C(1:2), cells should evolve to either C(2:2) or C(1:3) first, but neither genotype is commonly seen in these cells lines. In contrast, C(2:3) at 29.2% is the second most common genotype. If we include the two genotypes, C(2:4) and C(3:3), that are derivatives of C(2:3), this inclusive C(2:3) cluster is the most common genotype. The model of the next section helps to interpret the observation.

A model for the evolution of free-living cells

The pathways of chromosomal evolution can be diagrammed as a series steps in **Figure 5A**. Each node represents a C(Xa:A) genotype, the abundance of which is reflected in the size of the node. Thicker arrows indicate faster transitions which add/delete one X while the thinner arrow denotes the slower transition of adding/deleting the whole set of autosomes. The fitness of each genotype, W, is assumed to be determined by the Xa/A ratio. In general, one would expect the wild type (W1) to be the fittest genotype and we particularly wish to know whether that is indeed the case here.

We first model the evolution under strict neutrality where all nodes have the same fitness. For simplicity, genotypes are grouped into 3 clusters centering around the 3 dominant genotypes, C(1:2), C(2:2) and C(2:3), the frequencies of which are x1, x2 and x3, respectively. Each cluster consists of the dominant genotype as well as the less common ones adjacent to it (see **Figure 5A**). For instance,

bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

x2 is the sum of the frequencies of C(2:2) and C(3:2) and x1 is those of C(1:2), C(1:1) and half of C(1:3). The frequency of the last one, being adjacent to both C(1:2) and C(2:3), is split between the two clusters. Tallying up the numbers in **Figure 4D**, we obtain x1 = 0.41, x2 = 0.092 and x3 = 0.482 with a total of 0.984, excluding the marginal genotypes. The analysis below can be expanded to account for each genotype separately. The transitions between clusters are defined as follows:

$$x_1(T) \underset{au}{\overset{u}{\rightleftharpoons}} x_2(T) \underset{bv}{\overset{v}{\rightleftharpoons}} x_3(T)$$

where u and v are the transition rates and $x_i(T)$ is the frequency of cluster i at time T. Let X(T) be the vector of $[x_1(T), x_2(T), x_3(T)]$, expressed as

$$X(T) = X(0)\begin{bmatrix} 1-u & u & 0 \\ au & 1-au-v & v \\ 0 & bv & 1-bv \end{bmatrix}^T \quad (1)$$

When T>> 0,

$$[x_1(T), x_2(T), x_3(T)] \sim [ab, b, 1]\ /z \quad (2)$$

where z = ab + b +1. The genotype frequencies evolve toward the equilibrium, [ab, b, 1] /z, which depends on a and b, but not u and v. We posit that a > 1 and b > 1 because, as the chromosome number increases, the probability of chromosome gain/loss increases as well. By Eq. 2, x1(T) > x2(T) > x3(T) when T >> 0. In short, the relative frequency should be in the descending order of C(1:2), C(2:2) and C(2:3) if there is no fitness difference among genotypes. This predicted inequality at T>> 0 is very different from the observed trend.

Eq. 2 assumes that cell lines have been evolving long enough to approach this equilibrium. A more appropriate representation should be X(T) where T can reflect the time a cell line has been in culture. It is algebraically simpler if T is measured by the rate of chromosomal changes, u or v, rather than by the actual cell generation (Eq. 1, **Figure. 5B** and legends). We also assume u > v as u involves only the X but v involves the whole set of autosomes. With the initial condition of X(0) = [1,0,0], **Figure 5B** shows that the C(2:3) cluster approaches the equilibrium more slowly than the other two clusters. Therefore, the observed high frequency of the C(2:3) cluster (x3 = 0.482 vs. x1 = 0.41 and x2 = 0.092) is incompatible with a neutrally evolving model of chromosome numbers. The discrepancy is true at all time points and is more pronounced at smaller T's.

Rejecting the neutral evolution model, we now incorporate fitness differences into **Figure 4A** with W1 = 1 [for C(1:2) and C(2:4)], W2 = 1+s [for C(2:2)] and W3 = 1+t [for C(2:3)] where s and t can either be positive or negative. Here, we add a fourth genotype, C(2:4). In the supplement, we model 4 genotypes with x1 – x4 for the frequencies of C(1:2), C(2:2), C(2:3) and C(2:4) respectively. An expanded transition matrix is used to model selection, followed by a normalization step (Eq. S1). The solution in the form of $X(T) = X(0)\ M^T$ is given in Eq. S2 and the equilibrium X(T) is given in Eq. S3.

We are particularly interested in whether t> 0 in the 4-genotype model, i.e., whether C(2:3) has a higher fitness than the wild type, C(1:2). We observe that [x1, x2, x3, x4] = [0.374, 0.087, 0,292, 0.128] where x3 = 0.292 is more than 3 times higher than x2 =0.087 and is close to x1 =0.374. Eq. S3 shows that s<0 is necessary for x2 to be smaller than x3, and t>0 is necessary for x3 to be close to x1 (see Supplement). **Figure 5C** is an example in which s = -0.5 and t = 0.5. The equilibrium at T >> 0 is indeed close to the observed values.

bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In conclusion, it appears that the extant state in multicellular organisms of C(1:2) is not the fittest genotype for free-living mammalian cells. The observed genotypic distributions suggest that C(2:3) may have a higher fitness than the wild type, C(1:2).

Discussion

Free-living mammalian cells like all living things speed up the evolution when the environment changes. The practice of cell culturing, however, is to slow down the evolution in order to preserve cell lines' usefulness as proxies for the source tissues. Nevertheless, changes are inevitable and the evolution of sex chromosomes is but one example. It should be noted that cell lines derived from cancerous tissues and normal tissues are different in one important aspect. Cell lines derived from normal tissues generally do not undergo karyotypic changes at an appreciable rate[20–22]. They are therefore much less responsive to selection in cultured conditions that favor new karyotypes. Cancer cell lines, having been through more rounds of passages, have generally experienced stronger selection more frequently than normal cell lines.

Our observations suggest that the extant X:A relationship (C(1:2)) may not be optimal for free-living mammalian cells. The highest fitness peak, instead, appears to be closer to the karyotype of C(2:3) as free-living cells reproducibly evolve toward this new karyotype. The fitness peaks in free-living cells being different from that of the multi-cellular organisms is not unexpected. With many possible conflicts between individual cells and the community of cells (i.e., the organism), the interest of the community may lie in its ability to regulate the growth potential of its constituents. Free-living cells, on the other hand, are driven by selection to realize their individual proliferative capacity relative to other cells.

The convergence among these many cell lines to C(2:3) is unexpected in the context of cancer evolution. The TCGA project (reference) has shown that cancer evolution is a process of divergence, not convergence. Indeed, only 2 genes have been mutated in more than 10% of all cancer cases and tumors of the same tissue origin from two different patients may often share no mutated genes at all[5][23]. Therefore, the karyotypic convergence reported here is rather unusual.

We note that C(2:3) toward which cultured cells evolved happens to be the smallest possible increase in the X/A ratio from C(1:2). The higher fitness of C(2:3) than C(1:2) in free-living cells may lend new clues to the debate about the evolution of mammalian sex chromosomes[16][24]. With X-inactivation, it has been suggested that $E_{X/A}$ could have been reduced, or even halved[14][24]. The debate is about whether, and by how much, $E_{X/A}$ might have increased in evolution. Our observation that free-living cells continue to evolve toward C(2:3) raised the possibility that the evolutionary increase in $E_{X/A}$ has not been complete, in comparison with the ancestral $E_{X/A}$.

Finally, this study of cancerous cell lines may also have medical implications. The common view that tumorigenesis is an evolutionary phenomenon posits that individual cells in tumors evolve to enhance self-interest[3][4][25][26]. A corollary would be that tumorigenesis may have taken the first few steps toward uni-cellularity. This extended view is supported by many expression studies as well as the higher likelihood of obtaining cell lines from tumors than from normal tissues[2]. An alternative view, posits that tumors remain multi-cellular in organization[27] . These different views have been critically examined recently[5]. It is possible that cancer cells in vivo may have been gradually evolving toward a new optimum. In that case, cancer cells in men and women are converging in their sex chromosome evolution and become more efficient in proliferation in this new de-sexualized state.

bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure legends

Figure 1. Convergence in sex chromosomes in culture human cells.
(A) Percentage of lines with LOH (loss of heterozygosity). Each black dot represents an autosome and the red dot represents X and Y. LOH in male and female lines are separately displayed on the X and Y-axes. (B) Percentage of cell lines with either one or two Xa's in male lines. (C) Expression level of Xist in male cell lines, [Y] means with or without Y chromosome. The number of cell lines is on the top of each bar. (D) Expression ratios of X-linked genes between Xa[Y] and XaXa[Y] cell lines. Each grey dot represents a gene, and significant differences are indicated by black dots (t-test, $p<0.05$). (E) Percentage of cell lines with either one or two Xa's in female lines with whole X chromosome LOH. Female lines with partial X's or non-LOH are not included because it's ambiguous to assign the activation of X's. (F) Expression level of Xist in female cell lines of different X karyotypes. XaO (female lines with a single X), XaXa (female lines with isodisomy of X), XaXb (non-LOH female lines with heterozygous X's). The number of cell lines is on the top of each bar. All lines except XaXb have very low levels of XIST, suggesting active X's. In XaXb lines, the degree of X-inactivation is variable.

Figure 2. (A-B) Representative images of X chromosome FISH in the A549 cell line (A) with two Xs and HeLa (B) with three Xs. DNA is stained with DAPI (blue), and the X chromosome is labeled with Cy3 (red). (C-D) The distribution of the copy number of X's among cells from A549 (n = 343) and HeLa (n = 170).

Figure 3. Increasing of expression ratio of X versus autosome (EX/A)
(A) EX/A distributions among normal (N) and cancer (C) cell lines. NF and NM (or CF and CM) designate normal (or cancer) female and male lines. EX/A in cancer cell lines become larger than those of the normal cell lines. Note that the expression in normal cell lines is narrowly distributed and is close to that of the normal tissue when compared. Although the numbers of NF and NM lines are much smaller than CF and CM lines (17 and 24 vs. 279 and 341), their EX/A distributions are much tighter than in cancer cell lines. The actual counts correspond to kernel density are given in Fig. S4. (B) EX/A ratio in CF, CM, NF and NM lines with filtering with three different cutoffs (see methods). EX/A ratios are consistently higher in CF and CM lines than in NF and NM lines.

Figure 4. Autosomes change in a cohesive manner and coevolution of X and A.
(A) The density plot of autosome copy number among 620 cell lines shows peaks at 2 and 3 per autosome. (B-C) The observed distributions of gain in copy number among autosomes in two HeLa sublines. The expected Poisson distributions are also given for sublines with different means (λ = 10/13, 16/13; see text). (D) The percentages of C(Xa:A) types among the 620 cell lines.

Figure 5 – A model of karyotypic evolution driven by fitness differences.
(A) Evolutionary pathways of chromosomal changes. Each node represents a karyotype C(Xa:A) and the size roughly corresponds to its frequency. Cell fitness is assumed to be a function of the Xa/A ratio, which is represented by the Y-axis. The four abundant karyotypes are shown by solid black circles. Red arrows indicate faster changes in X and black arrows indicate slower changes in autosome. Main transitions between the common karyotypes are indicated by thicker arrows. (B) Changes in the frequencies of the three key genotypes as a function of time (T, expressed in units of 1/v) under fitness neutrality with all Wi's = 1. The parameters for Eqs. 1 and 2 are u = 10v, a =2 and b = 1.5. Both the theoretical trajectories and the observed values are given. The C(2:3) cluster (x3) is far more common in the observation than in the neutral model. (C) Changes in 4 karyotypic frequencies under selection according to Eq. S3 with s = -0,5 and t = 0.5. All other conditions are

bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

345 the same as above. Under selection, a reasonable agreement between the model and the observation
346 can be obtained.
347

bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Materials and Methods

Chromosome number estimation of HeLa sub-lines.

The processing of clonal expansion and whole genome sequencing of HeLa lines are described at Zhang et. al. (unpublished data). For each line, the copies of each chromosome are estimated according to the average sequencing depth by Control-FREEC, a tool for assessing copy number using next generation sequencing data[28].

Data collection

Three large-scale datasets were used in this study[8,29,30].

Genome-wide SNP array data on cancer cell lines and a normal training set were downloaded from The Wellcome Trust Sanger Institute under the data transfer agreement. Among the 755 cancer cell lines, 645 (from 288 females and 357 males) with available gender information were used for genotype information analysis in the present study. The processed data are in PICNIC output file format, which includes information on genotype, loss of heterozygosity and absolute allelic copy number segmentation[8]. Greenman et. al. developed the algorithm, PICNIC (Predicting Integral Copy Number In Cancer), to predict absolute allelic copy number variation in cancer[8]. This algorithm improved the normalization of the data and the determination of the underlying copy number of each segment. It has been used for Affymetrix genome-wide SNP6.0 data from 755 cancer cell lines, which were derived from 32 tissues. The Affymetrix Genome-Wide SNP Array 6.0 has 1.8 million genetic markers, including more than 900,000 single nucleotide polymorphism probes (SNP probes) and more than 900,000 probes for the detection of copy number variation (CN probes).

The genome-wide gene expression data for 947 human cancer cell lines from 36 tumor types were generated by Barretina et al[29]. as part of the cancer cell line Encyclopedia (CCLE) project using Affymetrix U133 plus 2.0 arrays and are available from the CCLE project website (CCLE_Expression_Entrez_2012-09-29.gct, http://www.broadinstitute.org/ccle/home). The expression profiles of 768 cell lines with gender information, representing 337 females and 431 males, were used in this study. These cell lines were partially overlapped with the lines used in Greenman et. al. Additionally, RNA-seq data from 41 lymphoblastoid cell lines from 17 females and 24 males were downloaded from GEO database (GSE16921)[30].

LOH detection and copy number estimation

We used the genotype information and absolute allelic copy number estimation generated from PICNIC to infer LOH, as well as copy number, of a specific chromosome. As for a chromosome, if ≥95% of SNP sites were homologous we considered that there was a LOH(loss of heterogeneity) event for this chromosome. Similarly, if ≥ 95% of detected alleles on the chromosome had a constant copy number of 0, 1, 2, 3 or 4, the copy number would be considered as the copy number of the chromosome. The copy number of the Y chromosome was estimated separately. In females, although all sites on Y chromosome should have yielded 0 copies, only ~ 60% of sites detected by the Y chromosome probes showed a copy number of 0. This result indicated that several X homologous regions on the Y were covered by ~30% of Y probes. Therefore, Y chromosome loss was defined as when more than 60% of SNP probes from the Y chromosome showed a copy number of 0.

Sex chromosome genotype inference

The expression level of *XIST* can be used as a proxy to distinguish the active X chromosome from the silent one as this gene was expressed on the inactive X chromosome and functioned *in cis*[13]. According to Greenman's and Barretina's studies, 496 cancer cell lines have both copy number and expression data. As expected, *XIST* was silenced in male cell lines,as well as in females with whole X chromosome LOH (Fig. 1C). Based on X chromosome LOH and copy number

bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

information, we identified five genotypes, including XaO (female lines with one X-20 lines), XaXa (female lines with isodisomy of X-17lines), XaXb (female lines with heterozygous for the X-28 lines), Xa[Y] (male lines with one X-53 lines) and XaXa[Y] (male lines with two X's-69 lines).

C(Xa:A)(ratio of active X's to autosomes) calculation

All male (341 lines) and female cell lines with whole X chromosome LOH (103 lines) were employed for C(Xa:A) calculation. C(Xa:A) was defined as the ratio of absolute X copy number to that of all autosomes.

$E_{X/A}$ (ratio of X to autosomal expression) calculation

$E_{X/A}$ was defined as the ratio of the expression of X-linked genes to that of autosomal ones. The median values of expressed X-linked and autosomal genes were used to calculate $E_{X/A}$ in both cancerous and normal cell lines. For the datasets from the Affymetrix U133 + 2.0 array, genes with signal intensities ≥ 32 ($\log_2$>5) were considered to be expressed. While as for RNA-seq data, genes with RPKM values ≥ 1 were considered to be expressed

Previous studies have shown that $E_{X/A}$ value may be affected by gene set used[15]. In addition, several silent genes in normal tissues have been shown to be expressed in tumor tissues[31]. Those genes were dominant on X chromosome, which could result in an increase of $E_{X/A.}$ To exclude the possibility that $E_{X/A}$ ratios may be biased in cancerous cell lines, gene sets for $E_{X/A}$ calculation were first selected in normal cell lines by three criteria, with the same sets then selected in cancerous cell lines. The three filtering criteria for gene set selection were RPKM >0, 1, and 5 in normal cell lines (Fig. 2*C*).

Differences in X-linked gene expression between Xa[Y] and XaXa[Y] lines

To explore the impact of extra X chromosome on gene expression levels of X-linked genes, 118 cell lines with Xa[Y] and 109 cell lines with XaXa[Y] were used. T-test with Benjamini and Hochberg adjusting method was employed to determine genes, the expression of which are significantly changed due to an extra X copy. 648 detected X-linked genes are plotted in Fig. 2*A*. The free statistical programming language R was used for the statistical analysis (version 3.0.1).

X chromosome Fluorescence *in situ* hybridization

HeLa cells (from the Culture Collection of the Chinese Academy of Sciences, Shanghai, China) were cultured in DMEM (Life Technologies) supplemented with 10% fetal bovine serum (FBS), 100 U/ml of penicillin, and 100 μg/ml of streptomycin. A549 cells (from Mi-lab) were cultured in RPMI-1640 (Life Technologies) with 10% fetal bovine serum (FBS), 100 U/ml of penicillin, and 100 μg/ml of streptomycin at 37℃ with 5% CO2. Approximately 2×10^6 cells were seeded and cultured in 10 cm dishes with 10 ml growth medium as described above. To synchronize the cells, 200 μl of thymidine (100 mM) was added to the cells. After incubating for 14 hours, the cells were washed twice with 10 ml PBS and then supplemented with 10 ml growth medium containing deoxycytidine (24 μM). After incubating for 2 hours, 10 μl nocodazole (100 μg/ml) was added to the cells. The cells were incubated for an additional 10 hours.

After synchronization, cells were harvested and treated with 4 ml hypotonic solution (75 mM, KCl) pre-warmed to 37°C for 30 min. The cells were then fixed via three immersions in fresh fixative solution (3:1 methanol:acetic acid) (15 min each time). The fixed cell suspension was spotted onto a clean microscope slide and allowed to air dry. We used the "XCyting Chromosome Paints" and "Xcyting Centromere Enumeration Probe" (MetaSystems, Germany) for whole X chromosomes and centromere of X chromosome fluorescence *in situ* hybridization (FISH) analysis, respectively. Following the manufacturer's instructions, 10 μl of probe mixture was added to the prepared slide. The slide was then covered with 22 x 22 mm^2 cover slip and sealed with rubber

bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cement. Next, the slide was heated at 75ºC for 2 min on a hotplate to denature the sample and probes simultaneously, followed by incubation in a humidified chamber at 37°C overnight for hybridization. After hybridization, the slide was washed in 0.4 x SSC (pH 7.0) at 72ºC for 2 min, then in 2 x SSC and 0.05% Tween-20 (pH 7.0) at room temperature for 30 seconds, before being rinsed briefly in distilled water to avoid crystal formation. The slide was drained and allowed to air dry. Finally, 5 µl DAPI (MetaSystems) was applied to the hybridization region and covered with a coverslip. The slide was processed and captured using fluorescence microscopy as recommended (Olympus FV1000, 100 x objective). The number of Xs were counted for each individual cell. A total of 343 HeLa cells and 170 A549 cells were screened.

Acknowledgments: We thank Jian Lu and Xionglei He for comments on an earlier version of the manuscript. Additionally, we thank Yang Shen, Yong E. Zhang, Rui Zhang and Wenfeng Qian for numerous constructive discussions.

Competing interests: The authors have declared that no competing interests exist.

bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Reference

1. Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and P.W. (2002). Isolating Cells and Growing Them in Culture. In Molecular Biology of the Cell (New York: Garland Science).
2. Hayflick, L. (1998). A Brief History of the Mortality and Immortality of Cultured Cells. Keio J. Med. *47*, 174–182.
3. Nowell, P. (1976). The clonal evolution of tumor cell populations. Science (80-.). *194*, 23–28.
4. Merlo, L.M.F., Pepper, J.W., Reid, B.J., and Maley, C.C. (2006). Cancer as an evolutionary and ecological process. Nat. Rev. Cancer *6*, 924–935.
5. Wu, C.-I., Wang, H.-Y., Ling, S., and Lu, X. (2016). The Ecology and Evolution of Cancer: The Ultra-Microevolutionary Process. Annu. Rev. Genet. *50*, annurev-genet-112414-054842.
6. Roschke, A. V., Tonon, G., Gehlhaus, K.S., McTyre, N., Bussey, K.J., Lababidi, S., Scudiero, D. a., Weinstein, J.N., and Kirsch, I.R. (2003). Karyotypic Complexity of the NCI-60 Drug-Screening Panel. Cancer Res. *63*, 8634–8647.
7. Charlesworth, B. (1991). Evoluton of Sex Chromosomes. Science (80-.). *251*, 1030–3.
8. Greenman, C.D., Bignell, G., Butler, A., Edkins, S., Hinton, J., Beare, D., Swamy, S., Santarius, T., Chen, L., Widaa, S., *et al.* (2010). PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. Biostatistics *11*, 164–75.
9. Guttenbach, M., Koschorz, B., Bernthaler, U., Grimm, T., and Schmid, M. (1995). Sex chromosome loss and aging: in situ hybridization studies on human interphase nuclei. Am. J. Hum. Genet. *57*, 1143–50.
10. Chow, J.C., Yen, Z., Ziesche, S.M., and Brown, C.J. (2005). Silencing of the mammalian X chromosome. Annu. Rev. Genomics Hum. Genet. *6*, 69–92.
11. Plath, K., Mlynarczyk-Evans, S., Nusinow, D. a, and Panning, B. (2002). Xist RNA and the mechanism of X chromosome inactivation. Annu. Rev. Genet. *36*, 233–78.
12. Ng, K., Pullirsch, D., Leeb, M., and Wutz, A. (2007). Xist and the order of silencing. EMBO Rep. *8*, 34–9.
13. Richardson, A.L., Wang, Z.C., De Nicolo, A., Lu, X., Brown, M., Miron, A., Liao, X., Iglehart, J.D., Livingston, D.M., and Ganesan, S. (2006). X chromosomal abnormalities in basal-like human breast cancer. Cancer Cell *9*, 121–32.
14. Xiong, Y., Chen, X., Chen, Z., Wang, X., Shi, S., Wang, X., Zhang, J., and He, X. (2010). RNA sequencing shows no dosage compensation of the active X-chromosome. Nat. Genet. *42*, 1043–7.
15. Deng, X., Hiatt, J.B., Nguyen, D.K., Ercan, S., Sturgill, D., Hillier, L.W., Schlesinger, F., Davis, C. a, Reinke, V.J., Gingeras, T.R., *et al.* (2011). Evidence for compensatory upregulation of expressed X-linked genes in mammals, Caenorhabditis elegans and Drosophila melanogaster. Nat. Genet. *43*, 1179–85.
16. Kharchenko, P. V, Xi, R., and Park, P.J. (2011). Evidence for dosage compensation between the X chromosome and autosomes in mammals. Nat. Genet. *43*, 1167-9–2.
17. Taylor, a I. (1968). Autosomal trisomy syndromes: a detailed study of 27 cases of Edwards' syndrome and 27 cases of Patau's syndrome. J. Med. Genet. *5*, 227–252.
18. Patterson, D. (2009). Molecular genetic analysis of Down syndrome. Hum. Genet. *126*, 195–214.
19. Kleijer, W.J., van der Sterre, M.L.T., Garritsen, V.H., Raams, A., and Jaspers, N.G.J. (2006). Prenatal diagnosis of the Cockayne syndrome: survey of 15 years experience. Prenat. Diagn. *26*, 980–984.
20. Shirley, M.D., Baugher, J.D., Stevens, E.L., Tang, Z., Gerry, N., Beiswanger, C.M., Berlin, D.S., and Pevsner, J. (2012). Chromosomal variation in lymphoblastoid cell lines. Hum. Mutat. *33*, 1075–1086.
21. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont,

bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

J.W., Boudreau, A., Hardenbol, P., Leal, S.M., *et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

22. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–72.
23. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., *et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.
24. Lin, F., Xing, K., Zhang, J., and He, X. (2012). Expression reduction in mammalian X chromosome evolution refutes Ohno ' s hypothesis of dosage compensation.
25. Chen, H., Lin, F., Xing, K., and He, X. (2015). The reverse evolution from multicellularity to unicellularity during carcinogenesis. Nat. Commun. *6*, 6367.
26. Chen, H., and He, X. (2016). The Convergent Cancer Evolution toward a Single Cellular Destination. Mol. Biol. Evol. *33*, 4–12.
27. Almendro, V., Marusyk, A., and Polyak, K. (2012). Cellular Heterogeneity and Molecular Evolution in Cancer. Annu. Rev. Pathol. Mech. Dis. *8*, 121023133009008.
28. Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2012). Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics *28*, 423–425.
29. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, a a, Kim, S., Wilson, C.J., Lehar, J., Kryukov, G. V, Sonkin, D., *et al.* (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.
30. Cheung, V.G., Nayak, R.R., Wang, I.X., Elwyn, S., Cousins, S.M., Morley, M., and Spielman, R.S. (2010). Polymorphic Cis- and Trans-Regulation of Human Gene Expression. PLoS Biol. *8*, e1000480.
31. Hofmann, O., Caballero, O.L., Stevenson, B.J., Chen, Y.-T., Cohen, T., Chua, R., Maher, C. a, Panji, S., Schaefer, U., Kruger, A., *et al.* (2008). Genome-wide analysis of cancer/testis gene expression. Proc. Natl. Acad. Sci. U. S. A. *105*, 20422–7.

Figure 1

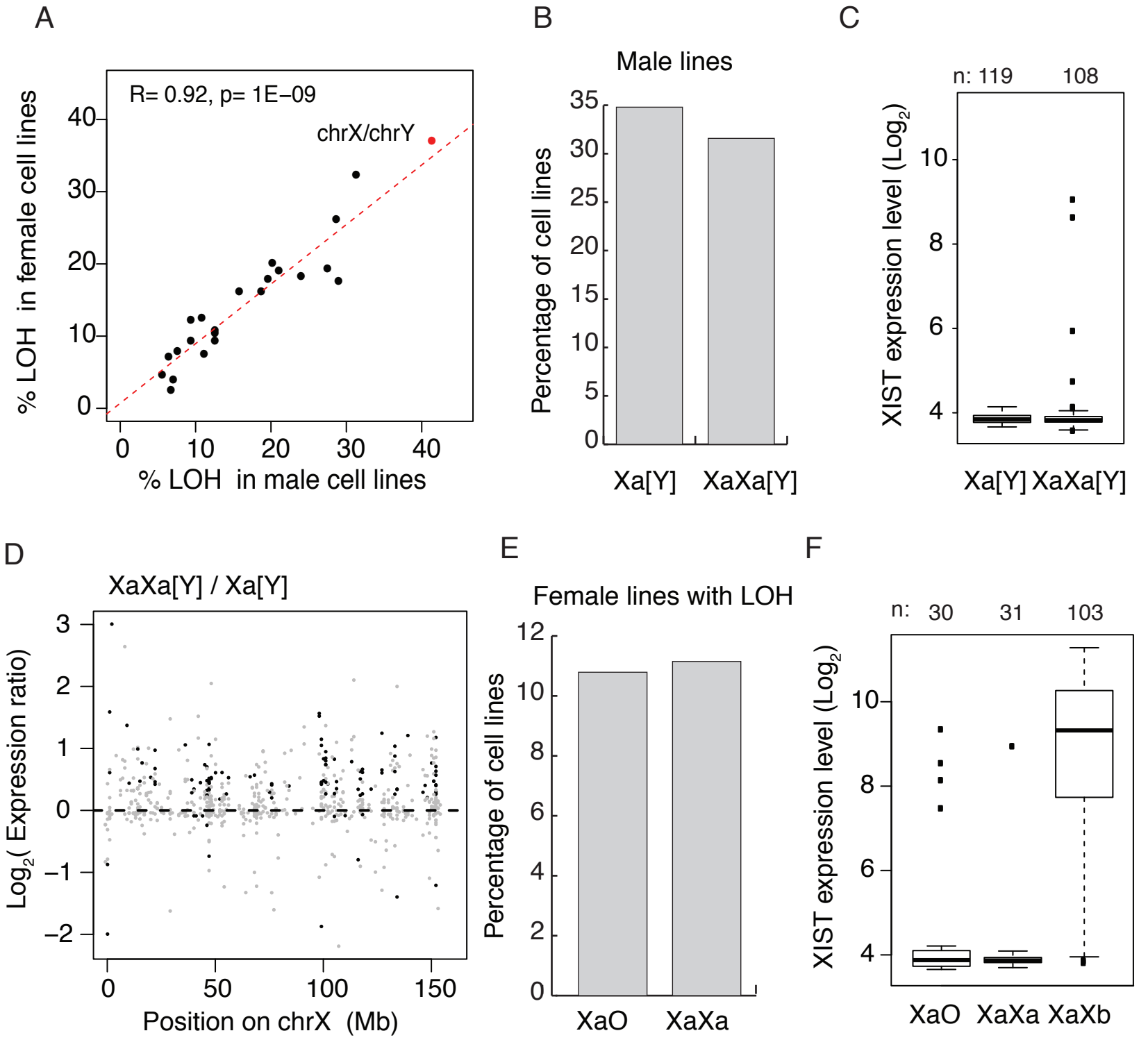

A
R= 0.92, p= 1E−09
chrX/chrY
% LOH in female cell lines
% LOH in male cell lines
B
Male lines
Percentage of cell lines
Xa[Y]
XaXa[Y]
C
n: 119
108
XIST expression level (Log2)
Xa[Y]
XaXa[Y]
D
XaXa[Y] / Xa[Y]
Log2(Expression ratio)
Position on chrX (Mb)
E
Female lines with LOH
Percentage of cell lines
XaO
XaXa
F
n: 30
31
103
XIST expression level (Log2)
XaO
XaXa
XaXb

Figure 2

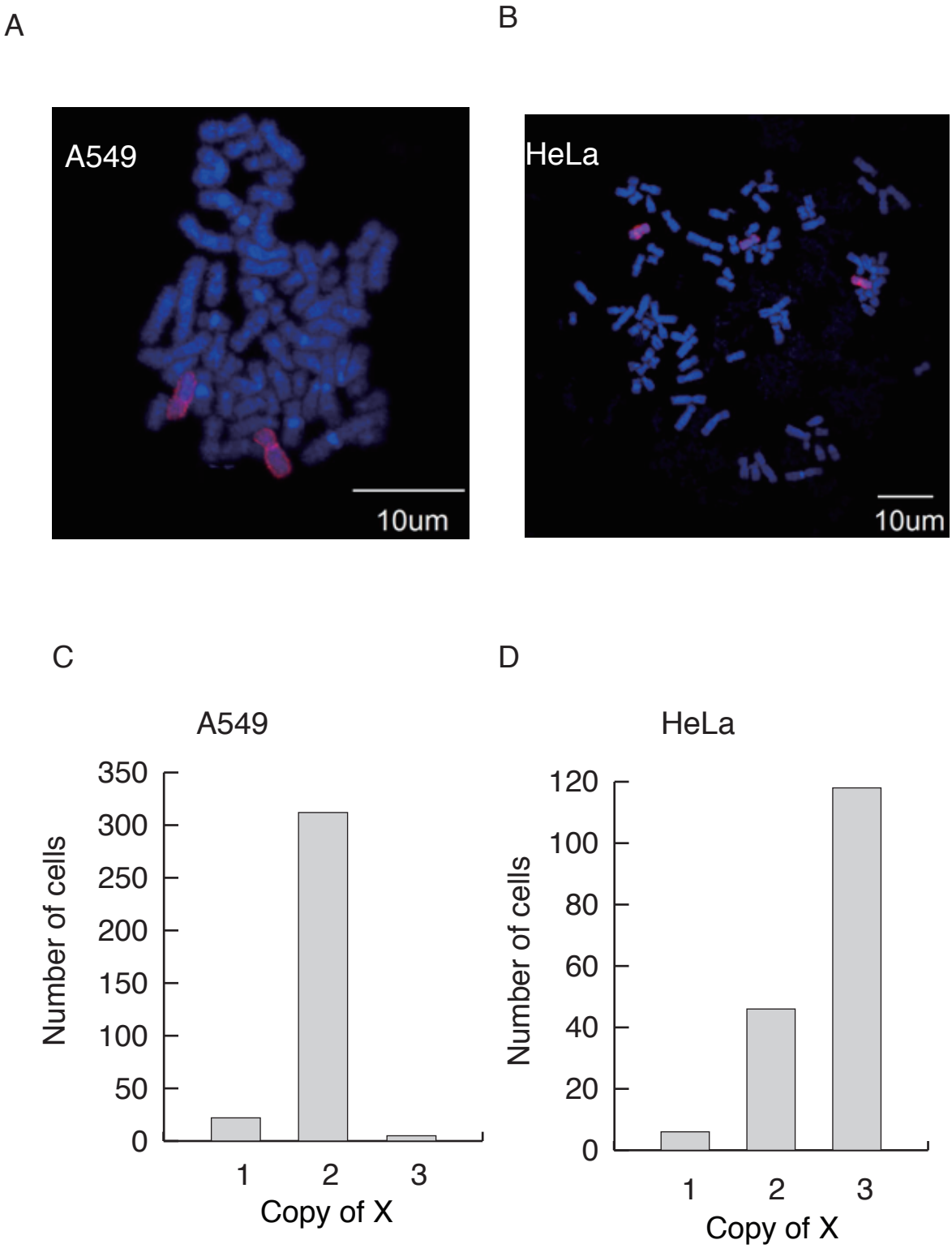

A
B
A549
10um
HeLa
10um
C
A549
Number of cells
350
300
250
200
150
100
50
0
1
2
3
Copy of X
D
HeLa
Number of cells
120
100
80
60
40
20
0
1
2
3
Copy of X

Figure 3

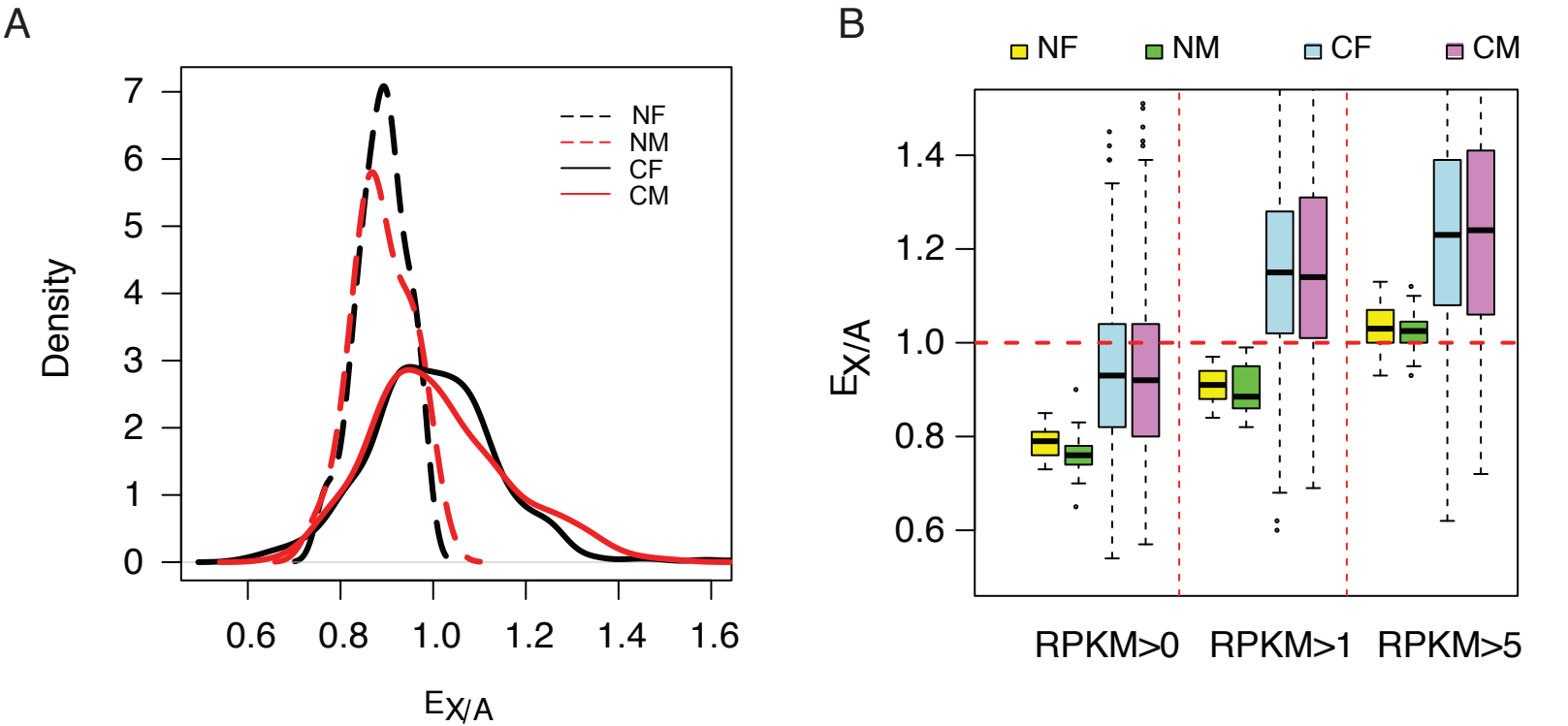
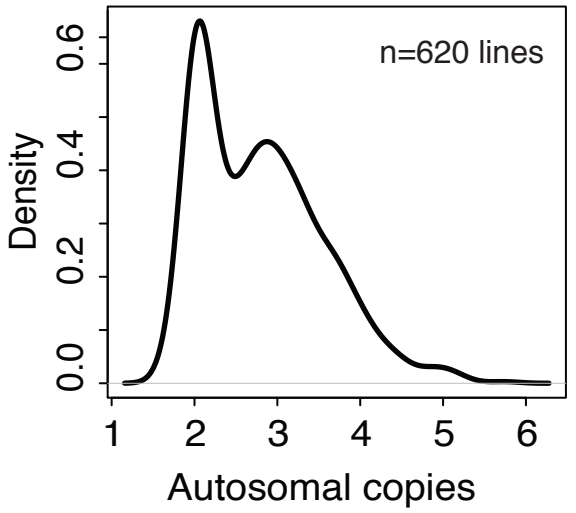
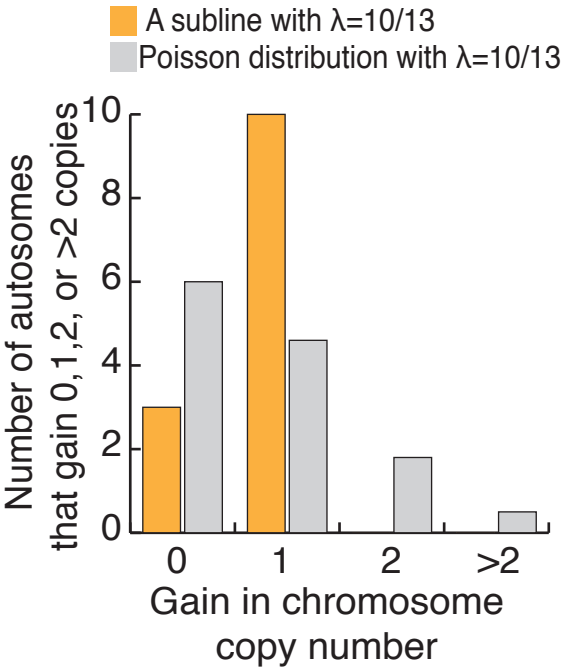

A
Density
NF
NM
CF
CM
0.6
0.8
1.0
1.2
1.4
1.6
EX/A
B
NF
NM
CF
CM
EX/A
RPKM>0
RPKM>1
RPKM>5

Figure 4

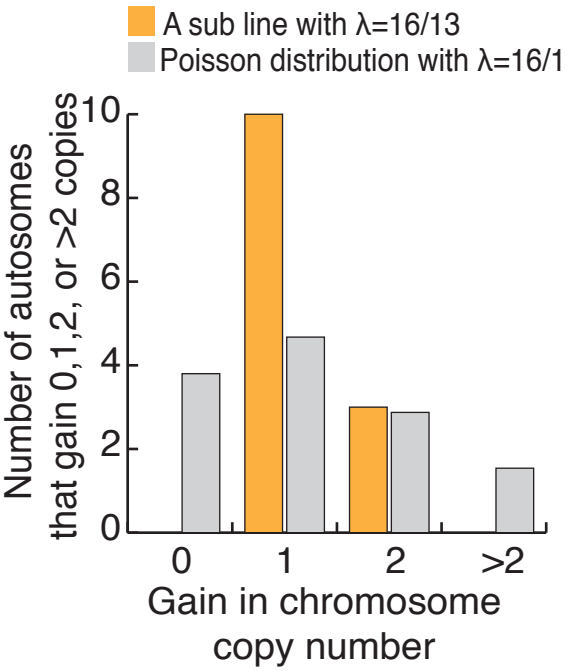
A



B



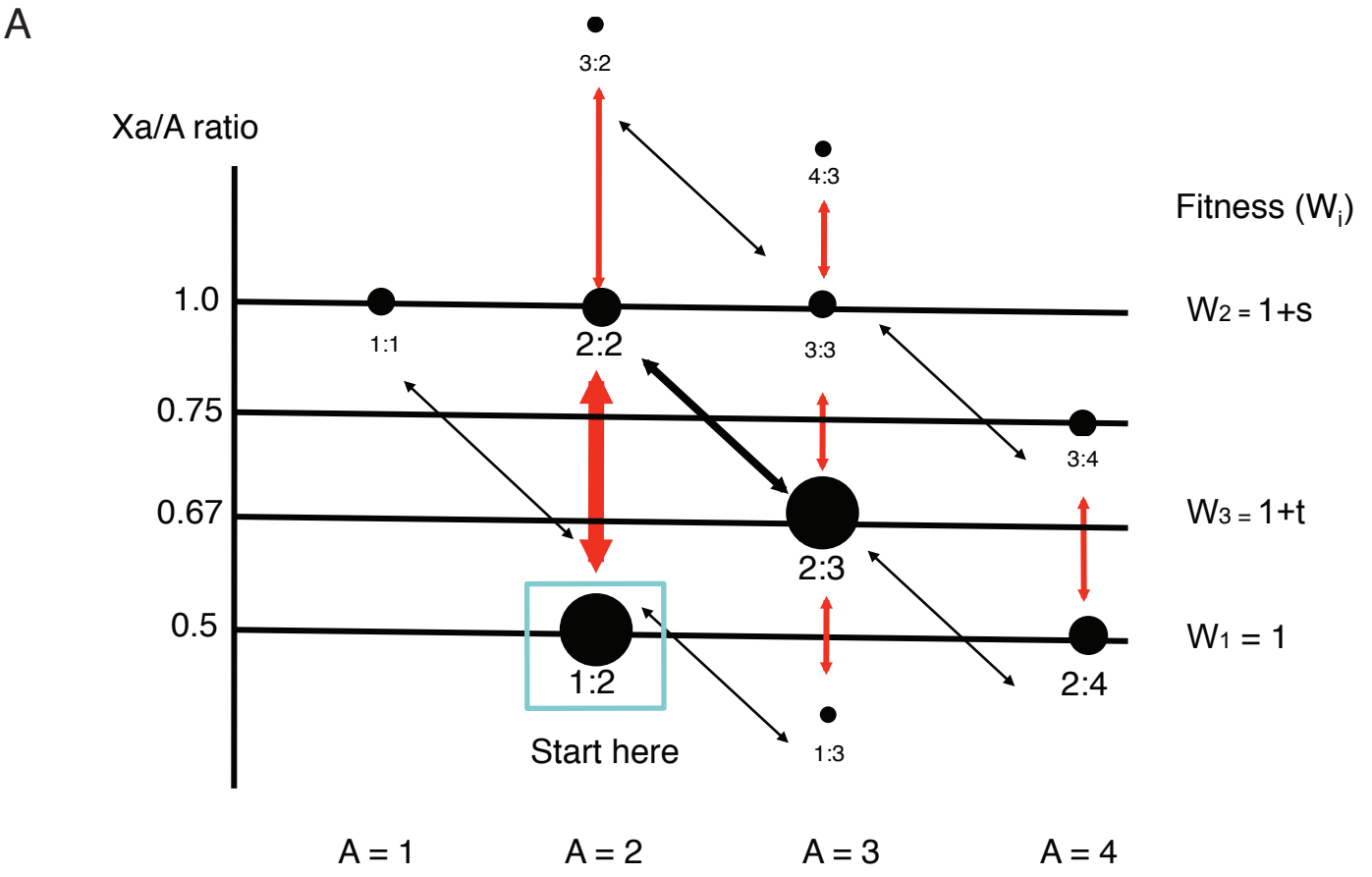
C

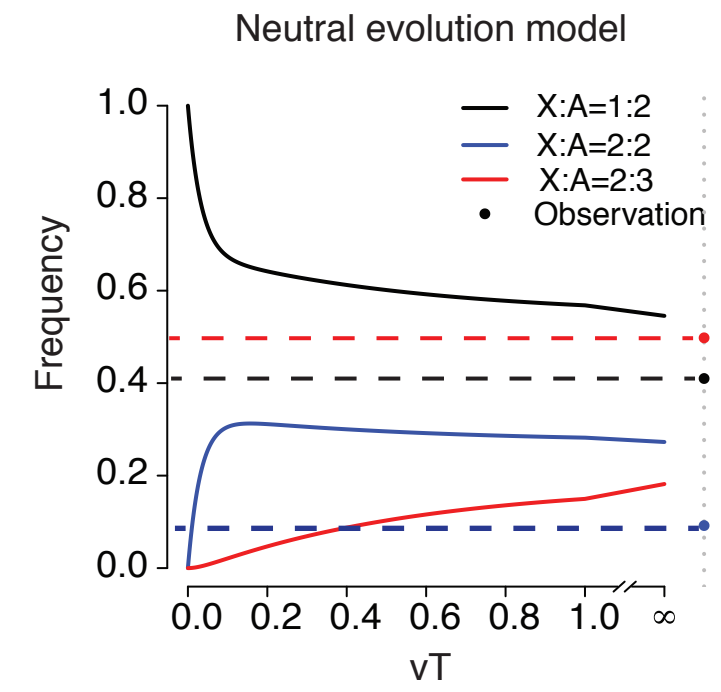


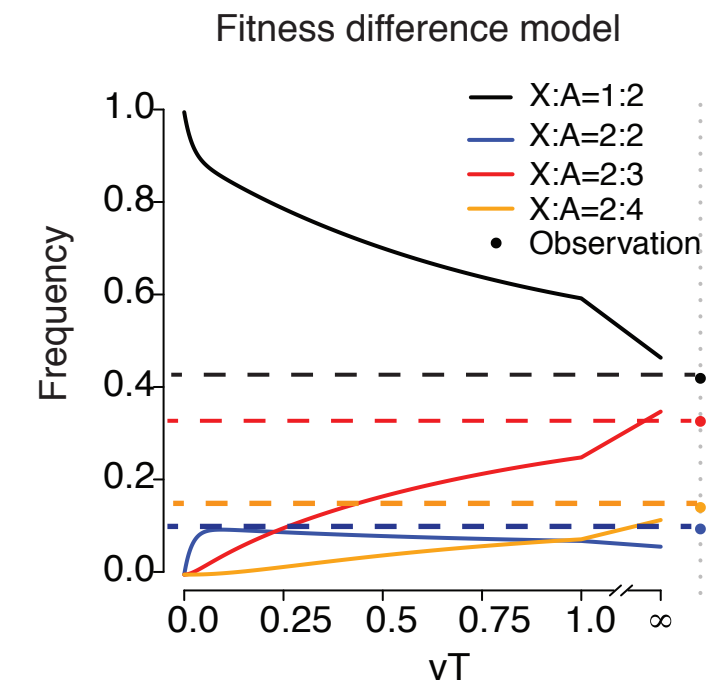
D

		Autosome			
		1	2	3	4
ChrX	1	0.0	37.4	7.3	0.0
	2	0.0	8.7	29.2	12.8
	3	0.0	0.5	2.5	1.1
	4	0.0	0.0	0.0	0.5

Figure 5


A
Xa/A ratio
3:2
4:3
Fitness (W_i)
1.0
1:1
2:2
3:3
$W_2 = 1+s$
0.75
3:4
0.67
2:3
$W_3 = 1+t$
0.5
1:2
$W_1 = 1$
2:4
Start here
1:3
A = 1
A = 2
A = 3
A = 4


B
Neutral evolution model
X:A=1:2
X:A=2:2
X:A=2:3
Observation
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.2 0.4 0.6 0.8 1.0 ∞
vT


C
Fitness difference model
X:A=1:2
X:A=2:2
X:A=2:3
X:A=2:4
Observation
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.25 0.5 0.75 1.0 ∞
vT

bioRxiv preprint doi: https://doi.org/10.1101/152223; this version posted June 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure S1: The frequency of chromosomes loss show negative correlation to their length. There is a slight tendency for the smaller autosomes to have higher LOH's than for the larger ones (R=~-0.4, p=~0.046). X chromosome shows significant deviated from the regression line.

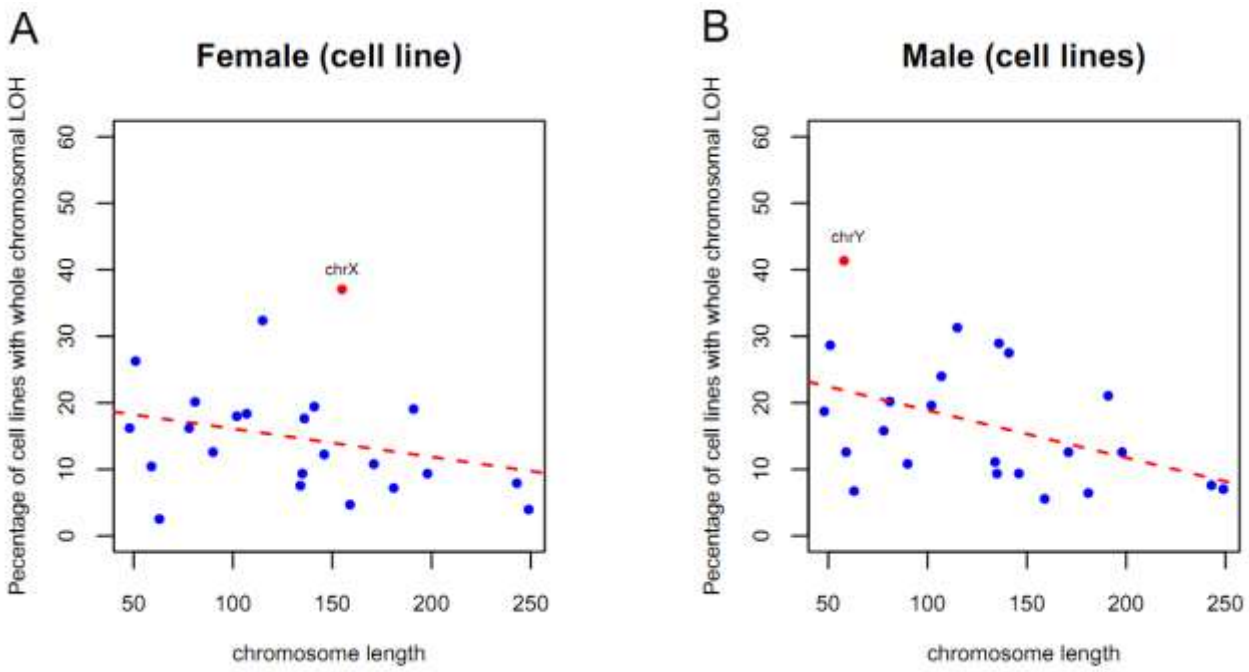


Figure S2: $E_{X/A}$ ratio in cancerous and normal cell lines by RNA-seq. The gene expression information (gtf files) by RNA-seq was downloaded from UCSC (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq/). There are 7 cancerous and 11 normal cell lines respectively. $E_{X/A}$ was calculated by the median value of expressed X-linked and autosomal genes. Expressed genes were selected as RPKM >1.

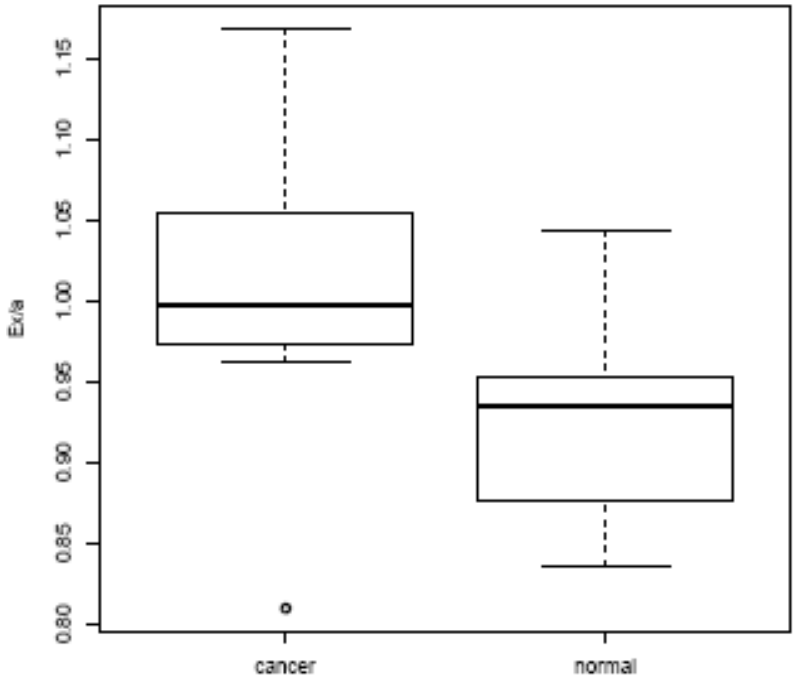


Figure S3: The frequency spectrum of $E_{X/A}$ in male and female cancerous cell lines compared to normal male and female cell lines. The median expression values of X and autosomes genes were used to compute $E_{X/A}$ for each cell lines. The proportion of cell lines within in a bin (0.1) was plotted as Y-axis. The $E_{X/A}$ of cancerous cell lines show a strong right shift compared to that of normal cell lines.

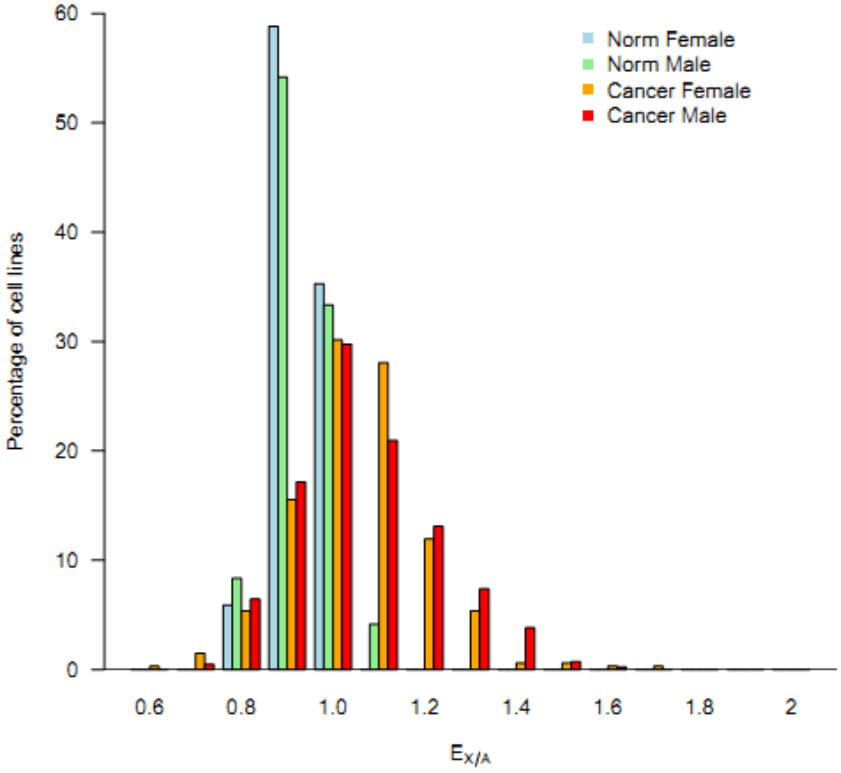


Figure S4: (A) Both the ancestral and sub-clonal HeLa population have 55-75 chromosomes centering around the triploid count of 69. (B) The composite distribution of the five lines with λ<1. And the comparison to Poisson distribution. (C)The frequency spectrum of C(Xa:A) in male and female cancerous cell lines. The median values for copy numbers on autosomes and X chromosome were used to calculate C(Xa:A). The proportion of cell lines within a bin (0.1) was plotted as the Y-axis. The discrete peaks denote the four major genotypes (X:3A; X:2A; 2X:3A; 2X:2A).