**The Underlying Sex Bias Genes of Sex Bias Epidemiology Pattern**

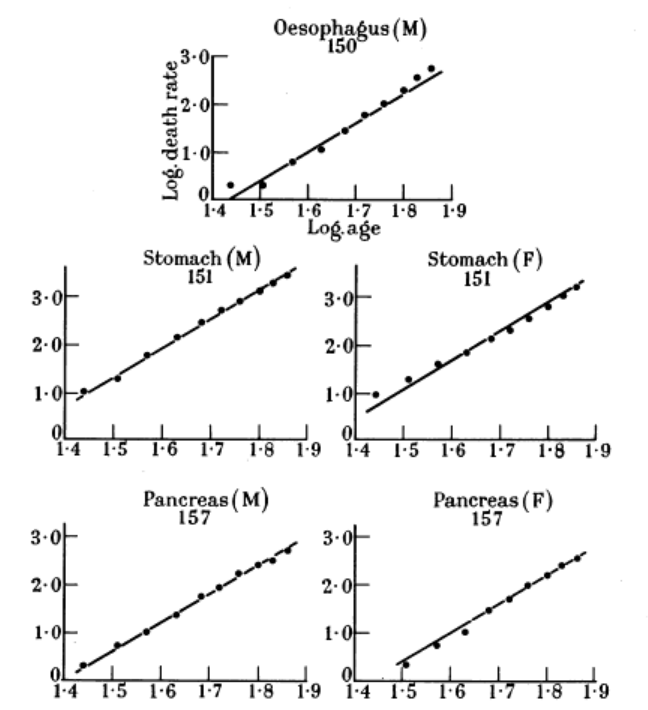
1. **研究背景**

癌症的进化发生受所处微环境的选择，同时具有一定的随机性，导致不同癌种，不同个体，乃至同一个体癌症组织内部的各个细胞，其突变谱不尽相同，这也是癌症异质性的来源。癌症的异质性对癌症的早期诊断及后期治疗都造成非常大的困难，如何应对癌症异质性十分迫切。

不同的种族具有不同的遗传背景。同一种族下，男女的基因组差异体现在性染色体，转录组差异则在各个器官均有所体现1。同一癌种，在这样的差异内环境下，肿瘤的进化发生是否会产生系统性的差异是一件值得研究的问题2。

1. **研究简介**

年龄是最大的致癌因子，其背后机理是人体细胞的更新需要干细胞分裂补充，而每一次干细胞分裂过程中的DNA复制都存在一定的突变概率。随着年龄的增大，干细胞DNA积累的突变增多，癌变风险随之增加。相关研究的数据很好展示了这一关系3。而实际上在上世纪50年代就已经有人观察到了这一现象并提出了著名的肿瘤发生多步模型（Multistage models of carcinogenesis）4–7。这一古老的模型至今仍在不断地被改进应用8–13，体现出强大的生命力。在肿瘤发生率和年龄的双对数图中（Fig 1），数据点被很好的线性回归，而其斜率被认为与hits number（即一个正常细胞变成癌细胞所需的必要突变次数）相关。

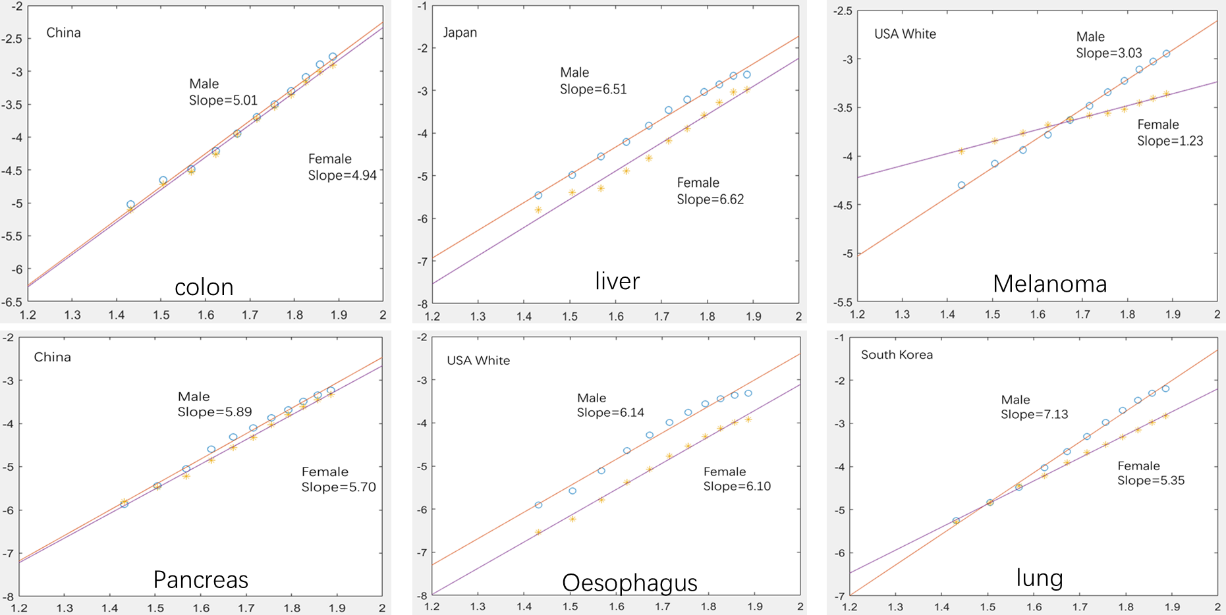


**Fig 1: Armitage和Doll 1954年文章中的癌症死亡率与年龄的双对数线性拟合图**5

我们使用世界卫生组织（WHO）的五大洲癌症数据（CI5-Xd）14，按人群（或种族）分组分析男女线型，发现男女发病率和年龄的双对数拟合直线线型在不同的人群（或种族）呈现出三种模式（Fig 2），分别为重合，平行分离，不平行。根据肿瘤发生多步模型，我们进一步将三种模式简化成两种，即平行和不平行。平行意味着hits number相同，不平行则意味着hits number不同。我们进一步猜测，hits number相同同时意味着该癌种癌变所需的driver genes也整体相同，而hits number不同则该癌种一定有男女差异突变的driver genes。

MD Anderson梁晗组针对性别差异做的pan-cancer分析15结果支持我们的假设，其揭示了某些癌种是存在性别差异突变基因和性别差异拷贝数变化基因的，而某些癌种则没有。不过该研究所覆盖的癌种并不全面，而且也只是简单地用统计方法抹平了人群（或种族）对两性差别的影响。

综上，本研究使用癌症发生的流行病学数据结合数学模型（Multistage models of carcinogenesis），试图探索同一人群（或种族）下癌症在性别层面上“异质性”及“同质性”。进一步用癌症基因组数据，定量检测是否存在“性别差异驱动基因”（sex bias driver gene），与模型预测结果相互印证。对于存在“性别差异驱动基因”的癌种，我们结合已有的研究对这种性别系统偏差给出假设。

****

**Fig 2: 男女发病率和年龄的双对数拟合直线线型的三种模式**

1. **检验流程**
2. **数据选取**
   1. 对于癌种的选择：
3. 因为需要比较性别差异，我们排除生殖类癌症。
4. 一般认为上皮类癌症（carcinoma）所在的组织在个体成熟后更新较为稳定,比较线性，符合multi-stage model，而儿童类癌症及液体癌症则不符合而被排除11。
5. 同时考虑癌症发病率统计的准确性（波动较小），和该癌种有在TCGA等数据库中的样本量，我们选取流行率较高的癌种。
6. 综上，所选择的癌种共11种（Table 1）。TCGA项目的癌种与WHO的统计类目里的癌种不完全对应（如TCGA的为Colon Adenocarcinoma，WHO只统计了部位Colon），我们比对的时候选择近似代替，因为TCGA的癌种均为该部位癌症的主要癌种，详见Table 1备注。
   1. 对于统计数据年龄段的选择：
7. 因为模型假设要求细胞更新均匀，我们排除成年前生长旺盛的阶段。这样同时可以排除很多具有家族遗传史的癌症高发人群，因为有遗传背景的人群倾向于较早发病。
8. 老年期细胞更新变慢，影响模型的线性关系。而且在年龄到达一定值之后，癌症发病率下降，普遍认为这部分老年人是经过选择，不易感癌症的人群。因此要将超出一定年龄的人群排除。
9. 综上两条，我们选择30-69岁年龄的发病数据。WHO数据每五岁一个统计区间，30-69岁分为30-34、35-39、40-44、45-49、50-54、55-59、60-64、64-69共 8个统计区间。

|  |  |  |
| --- | --- | --- |
| TCGA 项目癌种 | WHO统计癌种 | 备注 |
| Bladder Urothelial Carcinoma  (BLCA) | bladder transitional cell carcinoma | Urothelial carcinoma, also known as transitional cell carcinoma16 |
| Colon Adenocarcinoma  (COAD) | Colon | Adenocarcinoma is the most common form of colorectal cancer (>95%)17 |
| KIRC | kidney | KIRC representing approximately 92 percent18。 |
| Liver Hepatocellular Carcinoma  (LIHC) | Liver Hepatocellular Carcinoma |  |
| Lung Squamous cell carcinoma  (LUSC) | Lung Squamous cell carcinoma |  |
| Lung Adenocarcinoma  (LUAD) | Lung Adenocarcinoma |  |
| Melanoma  (SKCM) | Melanoma of skin |  |
| Esophageal Carcinoma  (ESCA) | Oesophagus Squamous cell carcinoma, Adenocarcinoma, Other specified carcinoma, Unspecified carcinoma | 将WHO中的四种Esophageal Carcinoma合并 |
| Pancreatic Adenocarcinoma  (PAAD) | Pancreas | over 80% are adenocarcinomas19 |
| Rectum Adenocarcinoma  (READ) | Rectum | Adenocarcinoma is the most common form of colorectal cancer (>95%)17 |
| Stomach Adenocarcinoma  (STAD) | Stomach | gastric adenocarcinoma comprises 95% of the total number of malignancies20 |

**Table 1 本次研究所选取的11个癌种及说明**

* 1. 对于WHO数据中人群（或种族）的选择：

这里的人群（或种族）指的是WHO数据中的register。选取的标准主要是我们对该人群的关注程度及数据质量。数据质量是指该register内的统计数据是否能真是体现当地癌症发病率，一般需要人口基数较大，且癌症诊断及统计工作比较完善（例如黑色人种我们选取美国的黑人数据而不是黑非洲国家的数据）。最为关心的人群是美国的白人人群，因为现在大部分可用癌症基因组数据来自这个群体。其次关心的为美国黑人人群，东亚三国（中国，韩国，日本）。本次初步分析选取的register全名单为： Canada、USA White、USA Black、South Korea、England、Scotland、Argentina、Brazil、China、Japan、India、France、Germany、Italy、Spain、Switzerland、Australia共17个人群（或种族）。

1. **定量化验证男女癌症发病率与年龄的双对数回归线是否平行**
   1. 不平行的定量化检验

我们选择17个人群（或种族）各11个癌种，共计17\*11=187组数据。每组数据男女各8个数据点共16个数据点。每组数据我们按照如下公式进行回归：

 Equation 1

可令原假设为β3=0，如检验显著推翻原假设则意味着男女各自对年龄的回归直线不平行。187组的P value我们使用bonferroni方法进行调整，以尽量减少假阳性的发生。

* 1. 平行的定量化检验

统计上证明两件事情相等是相对困难的，因此我们初试使用以下两种方法来进行判定，相互比较印证。

1. 等效性检验（Equivalence test）:

同样按Equation 1回归，考察β3的置信区间。如果β3的置信区间落在[-δ，δ]内，则认为β3等于0，即男女各自对年龄的回归直线平行。本次我们暂且将δ设为1。同时为了防止β3偏离0点，我们加入附加条件：



1. “CI检验（CI test）”：

我们将原来16个点男女分开各8个点拟合两条直线：



考察β1和β2的置信区间，如果满足下面的条件则认为两条线平行。



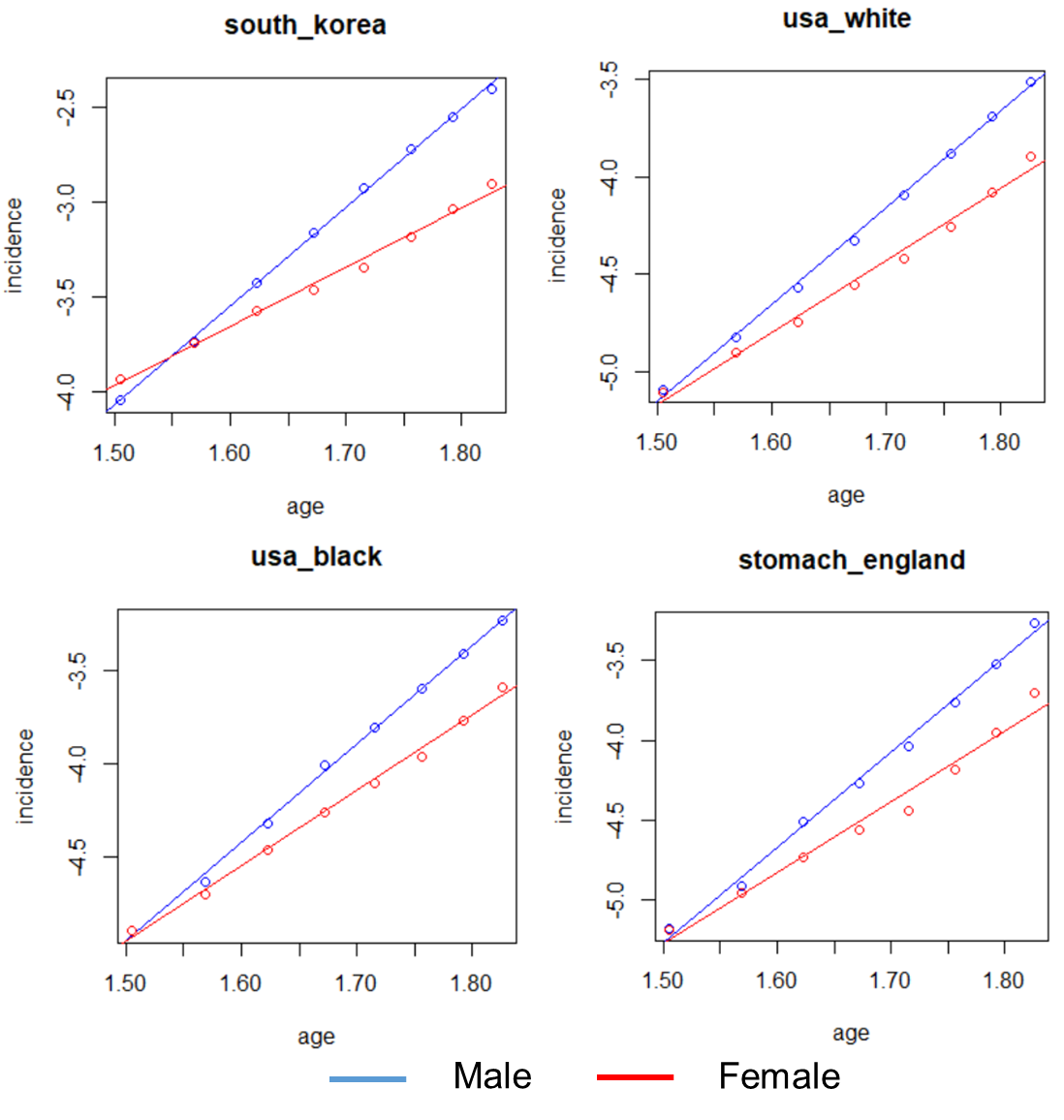
1. **在可用肿瘤基因组数据里检验是否有性别差异驱动基因（暂时只考虑突变和拷贝数变化两种形式）**
   1. 性别差异突变基因的检验

我们选用TCGA的MAF（Mutation Annotation Format）文件考察基因突变。流程如下：

1. TCGA每个癌种的MAF文件有4个，分别为4个pipeline（MuSE、Varscan 、SomaticSniper、MuTect）的结果，我们首先合并四个pipeline的结果确保抓到尽可能多的突变。
2. 剔除一些可能无意义的突变：Silent, Intron, IGR（Intergenic region）
3. 合并clinical数据（性别、种族等信息）
4. 筛选出美国白人的数据（目前只有该人群样本量充足）然后按性别分组
5. 统计出每组每个基因突变及没突变的样本量
6. 对每个基因在两组的突变频率是否存在差异进行卡方检验
7. 对所有基因检验的P value进行bonferroni调整
   1. 性别差异突拷贝数变化基因的检验

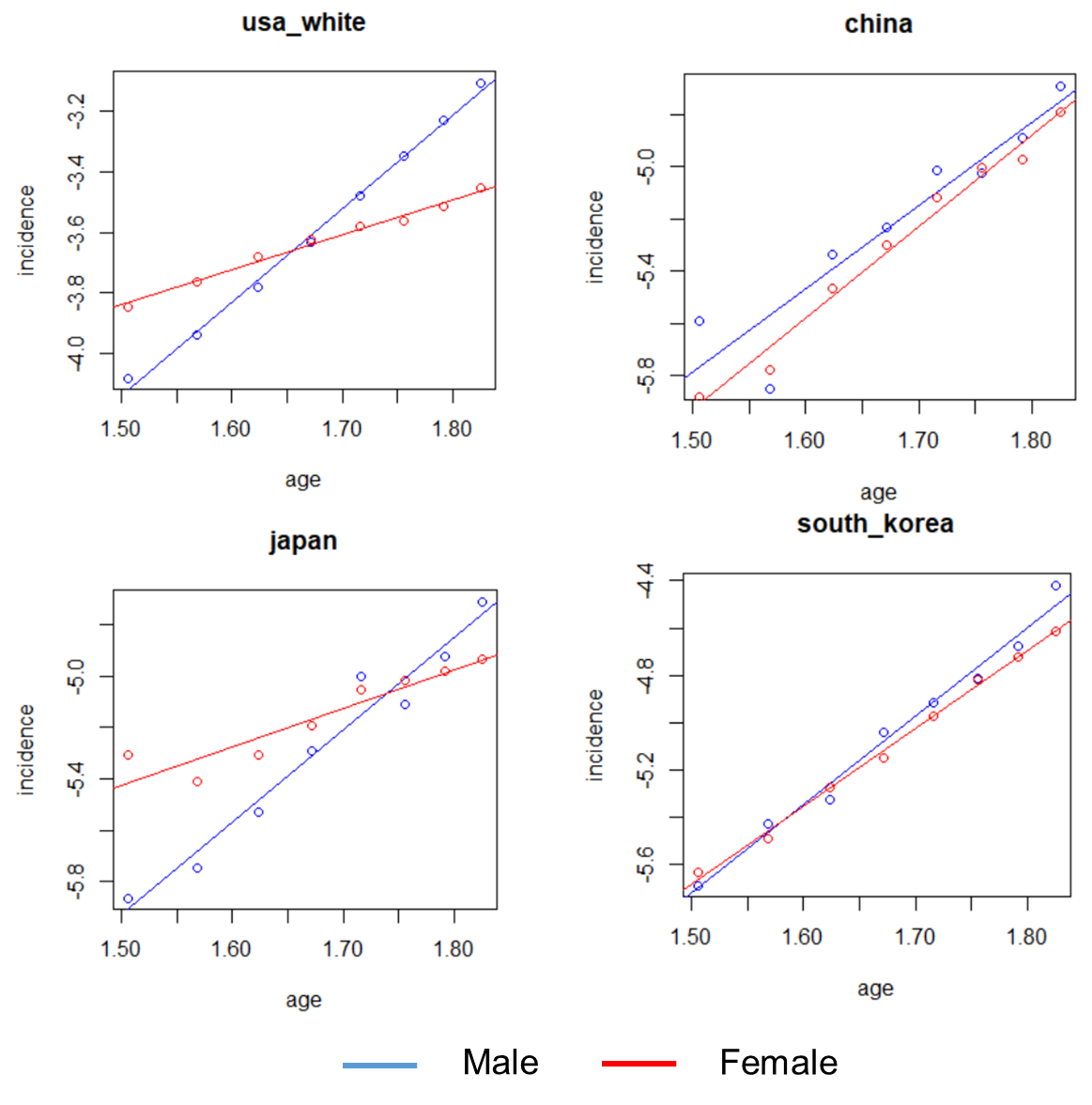
（暂未开展）

1. **检验结果**
2. **不平行的检验结果**
   1. 不平行模式最明显的为胃癌，除英格兰、阿根廷、印度和瑞士外，其余13个人群（或种族）全部检出不平行。



**Fig 3: 胃癌的典型模式图（韩国、美国白人和美国黑人检验为不平行，英国未被检出不平行）**

* 1. 其次为黑色素瘤，在加拿大、美国白人、英格兰、苏格兰、德国、意大利、西班牙和澳大利亚检出不平行，即除了瑞士、法国外，全部白人为主的人群（或地区）全部检出不平行。此外，日本检出不平行，而中国、韩国在下面的两种平行检验中分别被检验为平行。

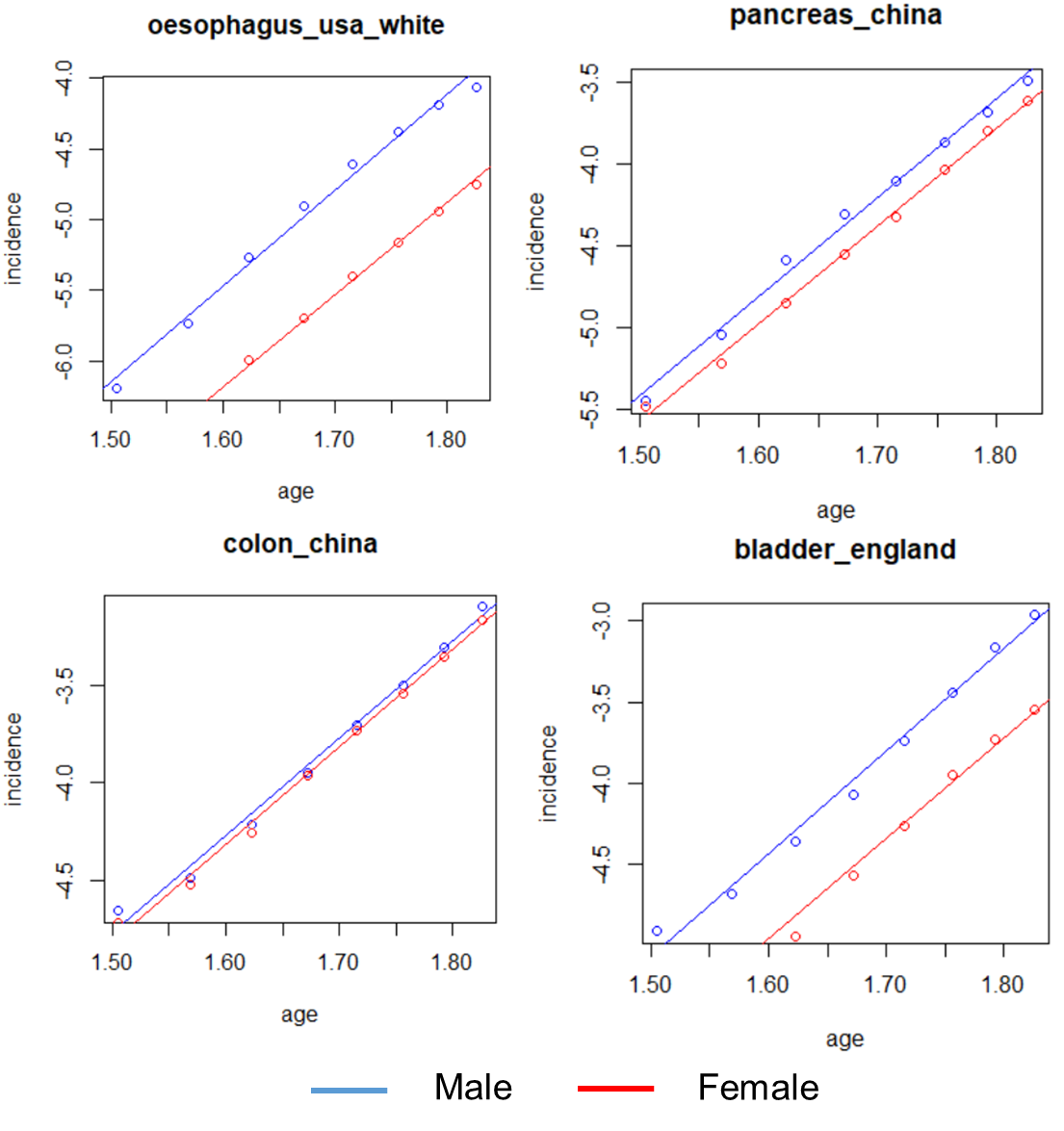


**Fig 4: 黑色素瘤的典型模式图（美国白人和日本检验为不平行，中国和韩国被两种平行检验分别检验出平行）**

* 1. 此外韩国人的两个肺癌亚种（LUAD、LUSC）被检出不平行，中国的LUSC被检出不平行，而日本未被检出。
  2. 其他还有一些癌种在个别地区被检出不平行，全部结果见附件result\_simple.csv。

1. **平行的检验结果**

等效性检验（Equivalence test）共检验出30个平行组，CI检验（CI test）共检验出42个平行组。两者交集为16组，并集为56组。全部结果见附件result\_simple.csv。

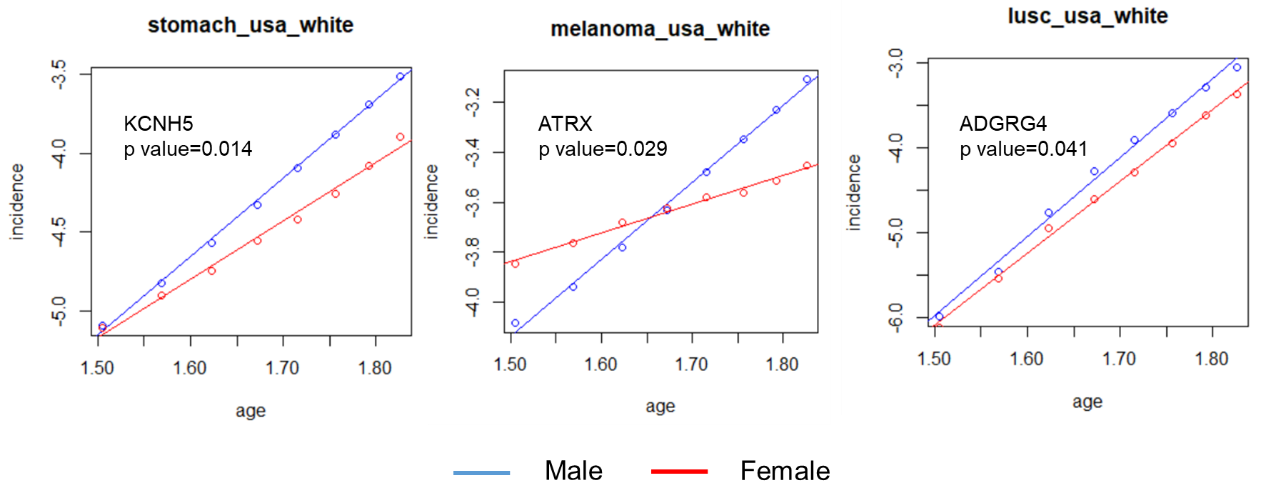


**Fig 5: 典型癌种的平行模式**

1. **性别差异突变基因的检验结果**

根据4.1的检验结果，美国白人在黑色素瘤、胃癌和肾癌被检验出不平行。此处我们在黑色素瘤检验出一个差异突变基因（KCNH5，p value=0.014），在胃癌检验出一个差异突变基因（ATRX，p value=0.029），在肾癌中未检出差异突变基因。此外我们在LUSC中检验出一个差别基因（ADGRG4，p value=0.041），而美国白人LUSC的不平行检验结果为不显著，也未被检验为平行。

对比梁晗组的结果15，其在肾癌中也未检验出差异突变基因（但有差异拷贝数变化基因），与我们的结果一致。其在LUAD检验出差异突变基因，而在LUSC中未检验出，这与我们的结果不一致，这可能是统计方法（调整种族，年龄等）造成的，需要进一步考虑。梁晗组未覆盖黑色素瘤与胃癌。



**Fig 6: 差异突变基因的检验结果与不平行检验结果对照**

1. **性别差异拷贝数变化基因的检验结果**

（暂未开展）

1. **差异基因review和机制假设**

（暂未开展）

1. **功能验证实验**

（暂未开展，视情况而定）

**Reference：**

1. Gershoni, M. & Pietrokovski, S. The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biol.* **15,** 7 (2017).

2. Clocchiatti, A., Cora, E., Zhang, Y. & Dotto, G. P. Sexual dimorphism in cancer. *Nat. Rev. Cancer* **16,** 330–339 (2016).

3. Tomasetti, C. & Vogelstein, B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347,** 78–81 (2015).

4. Nordling, C. O. A new theory on cancer-inducing mechanism. *Br. J. Cancer* **7,** 68–72 (1953).

5. Armitage, P. & Doll, R. The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis. *Br. J. Cancer* **8,** 1–12 (1954).

6. Armitage, P. Multistage models of carcinogenesis. *Environ. Health Perspect.* **63,** 195–201 (1985).

7. Armitage, P. & Doll, R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer* **91,** 1983–1989 (2004).

8. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.* **68,** 820–823 (1971).

9. Nunney, L. Lineage selection and the evolution of multistage carcinogenesis. *Proc. Biol. Sci.* **266,** 493–498 (1999).

10. Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G. & Vogelstein, B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 118–123 (2015).

11. Peto, R. Epidemiology, multistage models, and short-term mutagenicity tests. *Int. J. Epidemiol.* **45,** 621–637 (2016).

12. Moolgavkar, S. H. Commentary: Multistage carcinogenesis and epidemiological studies of cancer. *Int. J. Epidemiol.* **45,** 645–649 (2016).

13. Nowak, M. A. & Waclaw, B. Genes, environment, and ‘bad luck’. *Science* **355,** 1266–1267 (2017).

14. Forman D, Bray F, Brewster DH, Gombe Mbalawa C, Kohler B, Piñeros M, Steliarova-Foucher E, Swaminathan R and Ferlay J, editors (2013) Cancer Incidence in Five Continents, Vol. X (electronic version). Lyon: International Agency for Research on Cancer. Available from: http://ci5.iarc.fr, accessed [2017].

15. Yuan, Y. *et al.* Comprehensive Characterization of Molecular Differences in Cancer between Male and Female Patients. *Cancer Cell* **29,** 711–722 (2016).

16. What Is Bladder Cancer? Available at: https://www.cancer.org/cancer/bladder-cancer/about/what-is-bladder-cancer.html. (Accessed: 17th May 2018)

17. Thrumurthy, S. G., Thrumurthy, S. S. D., Gilbert, C. E., Ross, P. & Haji, A. Colorectal adenocarcinoma: risks, prevention and diagnosis. *BMJ* **354,** i3590 (2016).

18. Clear Cell Kidney Carcinoma. *The Cancer Genome Atlas - National Cancer Institute* Available at: https://cancergenome.nih.gov/cancersselected/kidneyclearcell. (Accessed: 17th May 2018)

19. Bond-Smith, G., Banga, N., Hammond, T. M. & Imber, C. J. Pancreatic adenocarcinoma. *BMJ* **344,** e2476 (2012).

20. Dicken, B. J. *et al.* Gastric Adenocarcinoma. *Ann. Surg.* **241,** 27–39 (2005).