使用深度神经网络预测人类突变的临床影响

nature 2018 July

cite:

Sundaram, L., Gao, H., Padigepati, S.R. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* **50,** 1161–1170 (2018). https://doi.org/10.1038/s41588-018-0167-z

摘要：由于疾病与基因突变之间的联系很难被区分出来，大量的测序数据的应用仍然有限。本文提出的框架在识别罕见病的致病突变上能够达到88%的准确率，并且在全基因组的意义上发现了14个影响智力发育的候选基因。从其他灵长类物种的常见突变能够提高对数以百万不确定其意义的突变的解释，进一步促进了人类基因测序的临床应用。

介绍：对于序列诊断的可解释性（interpretability）受限于难以解释的罕见基因突变。具有临床显著性的基因突变极其罕见，而且他们数量巨大，他们的影响尚未确定。数量大并且十分罕见的不确定临床意义的基因突变使得个体化用药和健康筛查构成了巨大的障碍。

大多数外显孟德尔病在人群中患病率非常低，这是在人群中观察到高频率的变异是良性结果的有力证据。分析不同人群中常见的突变是区分良性突变的有效策略，但是由于近代人类的一些瓶颈事件导致常见突变的数量丢失，导致常见突变的总量有限。Population

studies of present-day humans show a remarkable inflation from an effective population size (*N*e) of less than 10,000 individuals within the last 15,000–65,000 years, and the small pool of common polymorphisms traces back to the limited capacitance for variation in a population of this size. 在超过7000万的签字蛋白质改变的错义替换中，约由千分之一超过了总群等位基因频率0.1%。

黑猩猩与人类共享99.4的氨基酸序列。人和猩猩邻近关系暗示猩猩上的蛋白质突变的优化选择很可能对人类的相同状态的影响。

结果：**其灵长类物种的常见突变在人类上大都是良性的。**

聚合外显子组数据的最近可用性，包括外显子体聚集联盟收集了123,136人(ExAC)和基因组聚合数据库(gnomAD)，使我们能够测量自然选择对等位基因频谱中的错义突变和同义突变的影响。后面大概是说等位基因频率约低越可能是有害的突变。（有点看不懂）

个人理解：文中使用等位基因频率来描述基因的有害性，等位基因频率高时，错义突变与同义突变的比例低， 等位基因频率低时，错义突变与同义突变的比例高。

identical-by-state 状态一致的

PrimateAI:输入 human amino acid reference and alternate sequence centered at the variant, the position weight matrix conservation profiles calulcated from 99 vertebrate species(这部分好像并不能用于我们的序列先验)

pwm一直没弄清楚。