文献阅读：A high-resolution map of human evolutionary constraint using 29 mammals

元件注释

BGI-RD
深圳

October 19, 2011

Outline

研究背景

已知:

人类基因组中编码部分约 1.5%,HMRD 表明 5% 受到 purifying selection，其中又～ 3.5% 是非编码元件，可能与基因调控相关。

已有研究：

- HMRD:human-mouse-rat-dog
- Sipel:vertebrate

之前的比较基因组研究确定了这部分区域大体含量，但是不能检测到具体的 constraint element，因为分辨率不够，所以之前的工作只是针对 5% 中最保守的 top 5%。

本文工作

自 2005 年为 29 哺乳动物测序，以人类基因组做参考，确定高分辨率的 map，希望找到并细致分析保守元件。

- 肯定了之前关于保守元件约占总量 5% 的估计
- 确定了 4.2% 的序列是保守的，利用多种证据确定了其中 60% 是功能相关的
  - protein-coding
  - RNA
  - 调控区和 chromatin roles
- 提供了 exaptation 和加速进化的证据

Expation: 在进化过程中一些特征改变了最初的功能 -by Gould

http://en.wikipedia.org/wiki/Exaptation

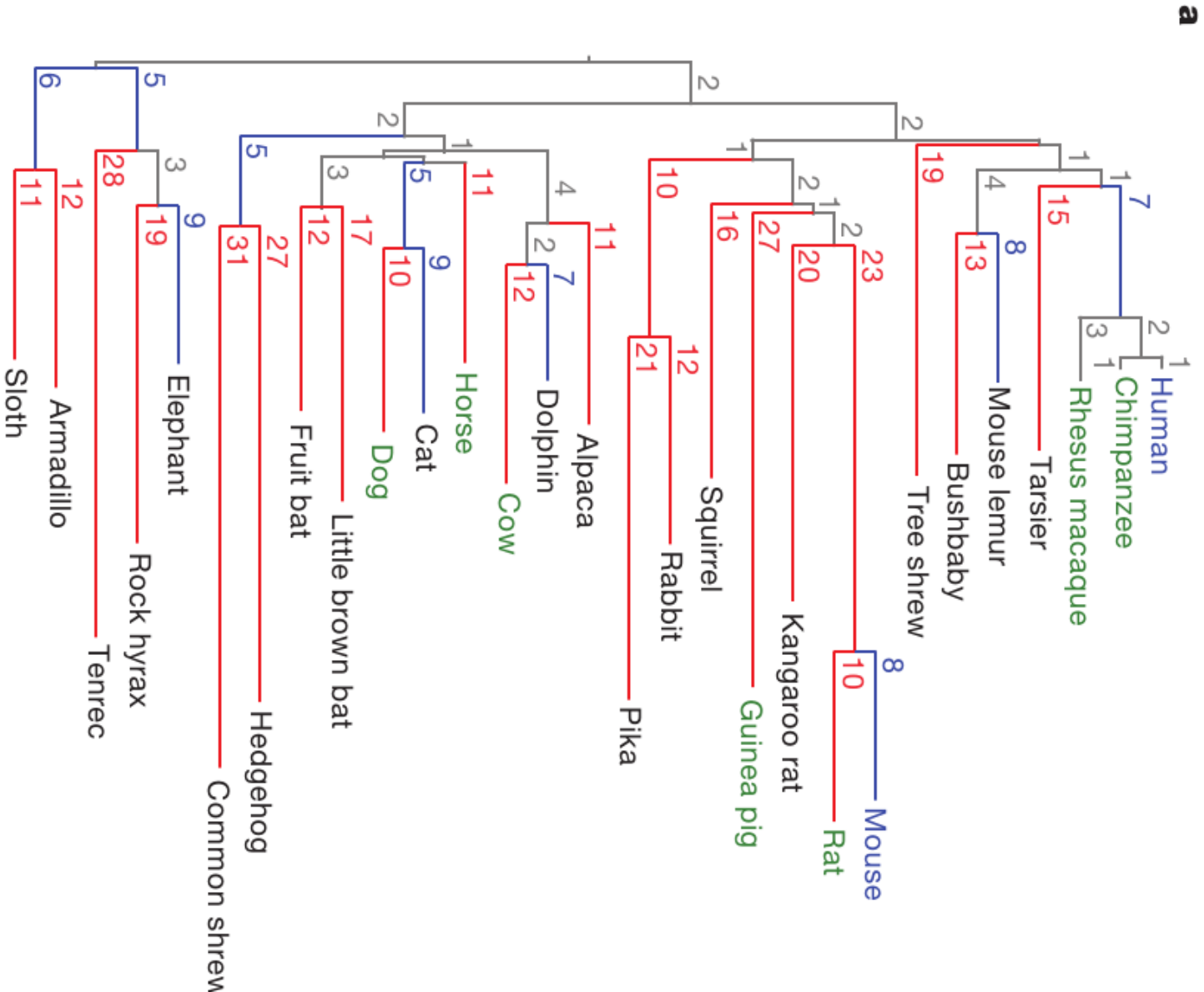
sequencing,asembly and alignment

- 7-fold:9 species previously described
- 2-fold:20 species first reported

20species 质量统计：

- contig size $N50_c$:2.8kb
- scaffold size $N50_s$:51.8kb
- 96% had quality score Q_{20},corresponding to a $<$1% error rate

sequencing,asembly and alignment



sequencing,asembly and alignment

Figure 1 | Phylogeny and constrained elements from the 29 eutherian mammalian genome sequences. **a**, A phylogenetic tree of all 29 mammals used in this analysis based on the substitution rates in the MultiZ alignments. Organisms with finished genome sequences are indicated in blue, high quality drafts in green and 2× assemblies in black. Substitutions per 100 bp are given for each branch; branches with ≥10 substitutions are coloured red, blue indicates <10 substitutions. **b**, At 10% FDR, 3.6 million constrained elements can be detected encompassing 4.2% of the genome, including a substantial fraction of newly detected bases (blue) compared to the union of the HMRD 50-bp + Siepel vertebrate elements[17] (see Supplementary Fig. 4b for comparison to HMRD elements only). The largest fraction of constraint can be seen in coding exons, introns and intergenic regions. For unique counts, the analysis was performed hierarchically: coding exons, 5′ UTRs, 3′ UTRs, promoters, pseudogenes, non-coding RNAs, introns, intergenic. The constrained bases are particularly enriched in coding transcripts and their promoters (Supplementary Fig. 4c).

sequencing,asembly and alignment

增加了物种间的差异，枝长显著增加。

- HMRD:0.68 substitution per site
- 20Ma:4.5 substitution per site

检验保守元件的能力主要取决于进化树的总枝长：the power to detect constrained elements depends largely on the total branch length of the phylogenetic tree connection the species.
Cooper,G.M, A quantitive estimates of sequence divergence for comparative analysis of mammaliang genomes

multiz 结果

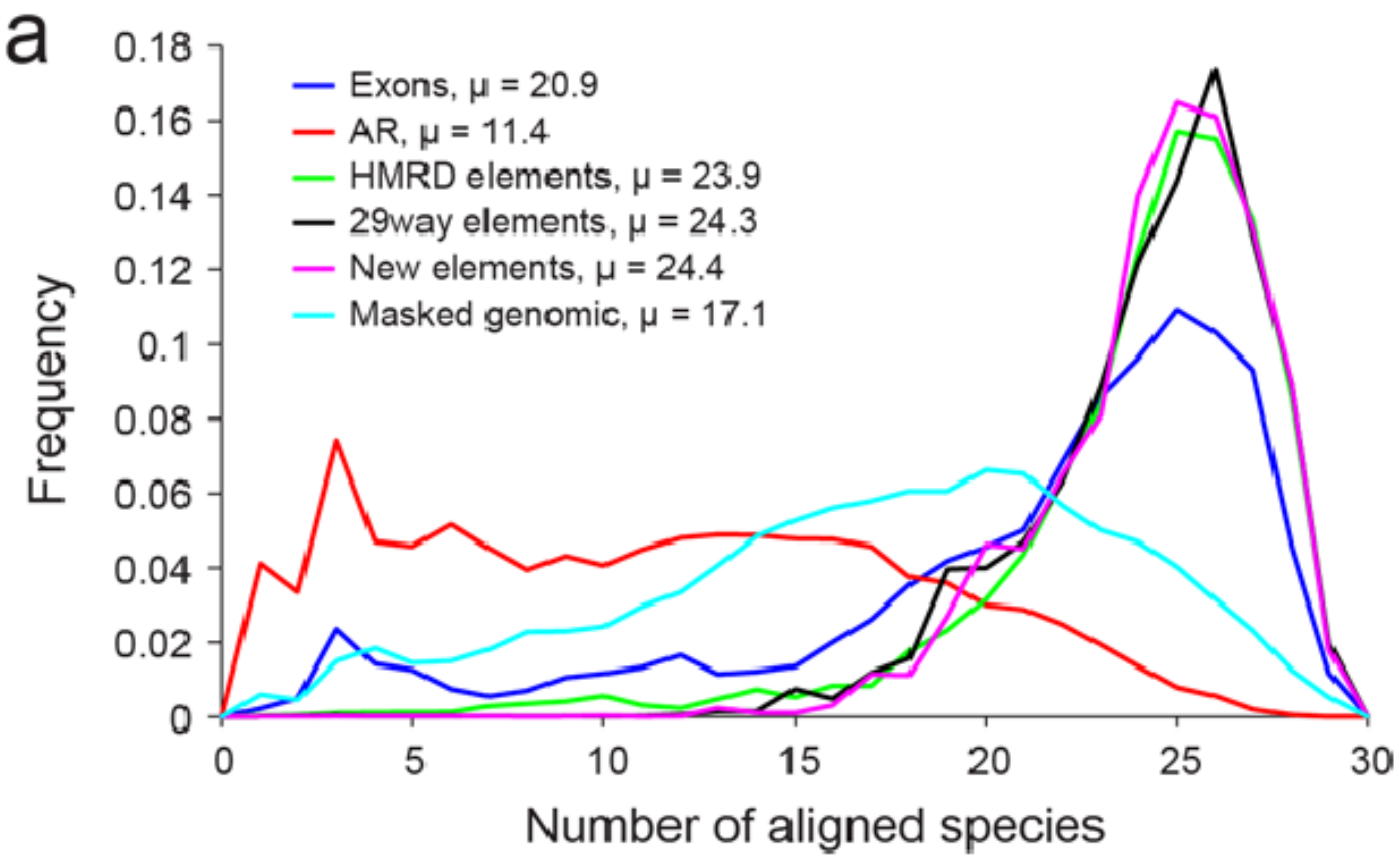
保守程度不同的元件类型有不同的比对层数
with a branch length of 4.3 substitutions per site:

- 20.9 at protein-coding positions in the human genomen
- 23.9 at the top 5% HMRD-conserved non-coding positions

with a branch length of 2.9 substitutions per site:

- 17.1 at whole-genome average
- 可能的原因是在对 whole-genome 做 multiz 时对非编码区的 large deletion ？？？这里需要看补充材料
- 11.4 at ancestral repeats
- 与非功能区域的重复序列相同

multiz 结果

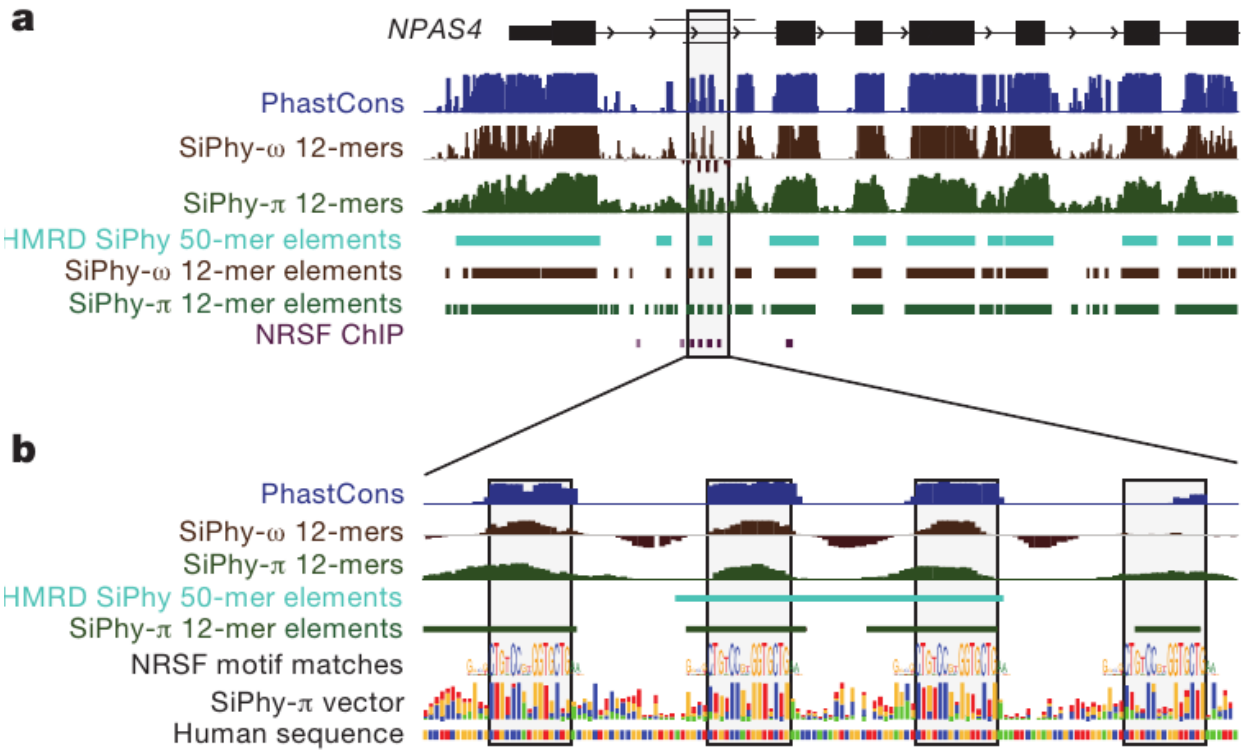

a
Frequency
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Exons, μ = 20.9
AR, μ = 11.4
HMRD elements, μ = 23.9
29way elements, μ = 24.3
New elements, μ = 24.4
Masked genomic, μ = 17.1
0
5
10
15
20
25
30
Number of aligned species

检测 constrained sequence

在整个基因组水平，用 SiPhy 方法估计，保守区域约 5.36% at 50bp wondows 5.44 at 12bp windows。平均长度 36bp，HMRD 平均长度 123bp，Sipel 平均长度 104bp。(图)
可以确定的：

- 29M:
  - 3.6 million elements
  - 4.2% at resolution of 12bp
- HMRD:
  - $< 0.1\%$ at resolution of 12bp
  - 0.2% at resolution of 50bp
- Sipel(5vertebrates)：
  - 4.1% 被检测到
  - 其中只有 45% 与 29M 一致，说明这次用 29M 检测到的很多是在哺乳动物中特有的。

检测 constrained sequence



29M 可以实现更高分辨率的保守元件检测。图中：NRSF 是神经元限制沉默因子，位于 NPAS4 基因的启动区。之前因为对短序列的检测能力不够，没有获得完整的结构。

碱基	T	C	A	G
T	0	1	1.5	1.5
C	1	0	1	1.5
A	1.5	1	0	1
G	1.5	1.5	1	0

谢谢