Reconstructing Duplication Histories of Complex Tandem Arrays in Multiple Primate Species

Tomáš Vinař[1], Giltae Song[3], Yu Zhang[3,4], NISC Comparative Sequencing Project[2], Eric D. Green[2], Webb Miller[3], Adam Siepel[1]

[1] Biological Statistics and Computational Biology, Cornell University, Ithaca, NY
[2] National Human Genome Research Institute, NIH, Bethesda, MD
[3] Center for Comparative Genomics and Bioinformatics, Penn State University, University Park, PA
[4] Department of Statistics, Penn State University, University Park, PA

Complex tandem arrays (CTAs) are regions of the genome that have evolved by series of segmental duplications local to particular genomic loci. They cover approximately 5% of the human genome, contain gene families linked to diseases such as various cancers, Alzheimer's, and HIV, and are strongly enriched for genes under positive selection. We are collaborating on an in-depth study of 14 biologically significant CTAs covering approximately 7 Mbp of the human genome. The study involves sequencing these regions in up to seven primates to overcome assembly problems and acquire data in novel organisms, development of comparative methods to reconstruct their duplication histories, and functional and positive selection analysis of the genes located in these regions.

Cluster selection: (finished)

- estimated cluster activity in various epochs of primate evolution from human sequence,
- complemented these estimates with array CGH data, comparing copy numbers of genes of several primate species (Dumas et al., 2007),
- selected final candidates, taking into account biomedical significance (gene function, disease associations).

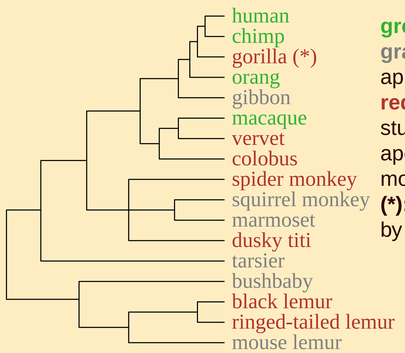
Sequencing project: (in progress)

- design probes targeting selected regions that work among all primates,
- screen the BAC libraries,
- Sanger sequencing of individual BACs followed by careful assembly.

Computational analysis: (methods development in progress)

- design algorithms to reconstruct duplication histories of individual clusters,
- use data from multiple species to establish direction and ordering of individual duplication events (insufficient information in single species),
- reconstruct multiple alignments of duplicated segments and corresponding segment trees in order to apply comparative genomics methods involving models of sequence evolution,
- examine bearing of duplication events on known functional sites and on disease-related genes,
- positive selection analysis.

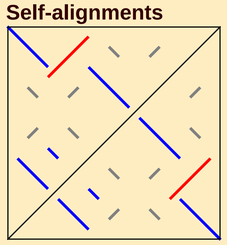
Choice of species for sequencing



green: whole genome sequenced
gray: whole genome in progress or approved (NHGRI, 2008)
red: additional species chosen in this study to span primate evolution (great ape, 2 old world monkeys, 2 new world monkeys, 2 prosimians)
(*): gorilla currently being sequenced by Sanger

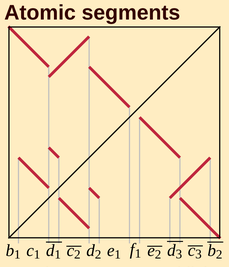
Estimating cluster dynamics (Zhang et al., 2008)

Self-alignments



Dot-plot after 3 events. Color indicates duplication age (from the most recent: red, blue, gray)

Atomic segments



Atomic segments establish reasonable abstraction and allow precise definition of a combinatorial problem.

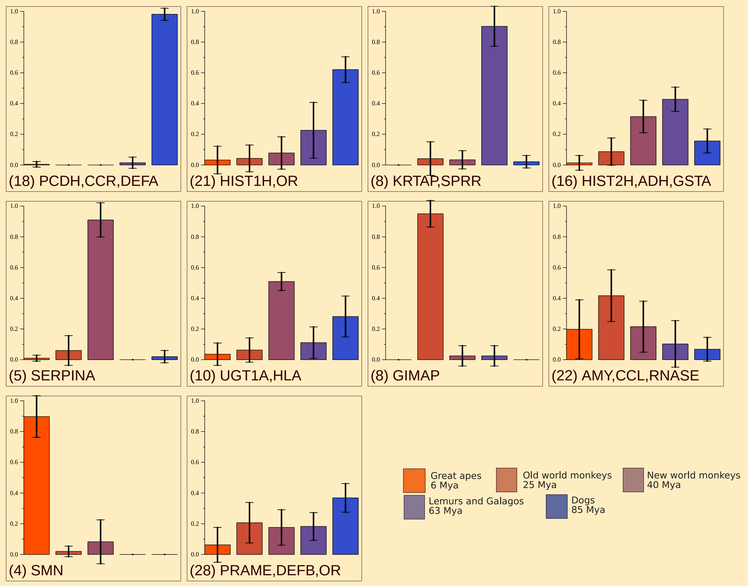
Reconstruction (in reverse)

$$b_1 c_1 \overline{d_1} \overline{c_2} d_2 e_1 f \overline{e_2} \overline{d_3} \overline{c_3} \overline{b_2}$$
$$\Uparrow \quad (\overline{d_3}\overline{c_3}, \overline{d_1}\overline{c_2})$$
$$b_1 c_1 d_2 e_1 f \overline{e_2} \overline{d_3} \overline{c_3} \overline{b_2}$$
$$\Uparrow \quad (b_1 c_1 d_2 e_1, \overline{e_2} \overline{d_3} \overline{c_3} \overline{b_2})$$
$$bcdef$$

- **Candidate alignments:** a simple **necessary condition** for an alignment to be the latest duplication.
- If **no breakpoint reuse,** every sequence of candidate alignments leads to an equal number of duplications $\Rightarrow$ **ambiguity**
- **Practical solution:** sequential importance sampling algorithm (also allows breakpoint reuse and deletions)
- Explore many histories, summarize their characteristics

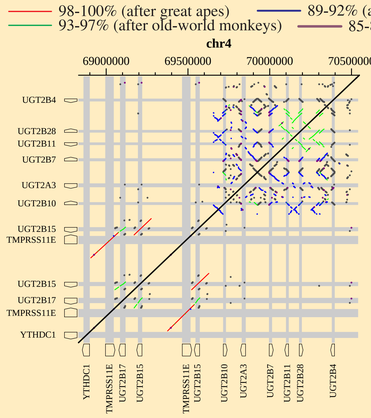
Distribution of predicted duplications over time

Clustering of 140 human CTAs based on distribution of duplication and deletion events in individual epochs. Legend of each profile shows number of CTAs and examples of CTAs in each category.



PCDH: protocadherin subfamily, neural cell adhesion. **CCR:** chemokine receptors. **DEFA,DEFB:** defensins (fungicide, antiviral, antibacterial). **HIST:** histones. **KRTAP:** keratines. **OR:** olfactory receptors. **GIMAP:** immuno-associated nucleotide-binding proteins. **AMY:** amylases, digestion of starch and glycogen. **SMN:** survival of motor neuron. **RNASE:** nonsecretory ribonuclease precursor. **PRAME:** preferentially expressed antigens in melanoma.

CTAs selected for sequencing



UDP-glycosyltransferases (UGT2)

Function: elimination of potential toxic xenobiotics

Diseases: prostate cancer, Criggler-Najjar syndrome

Total	Event distribution				
Events	GA	OWM	NWM	PROS	DOG
82	0.02	**0.30**	0.29	0.10	0.28

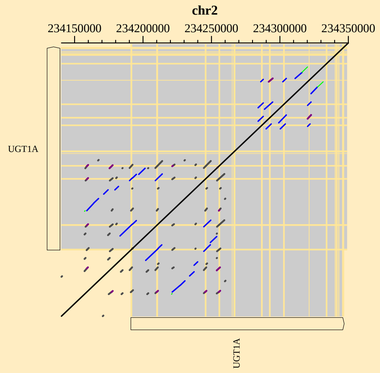
GA: after great apes; OWM: after old-world monkeys; NWM: after new-world monkeys; PROS: after prosimians; DOG: after dogs

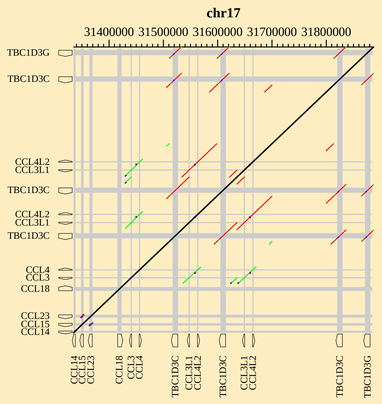
UDP-glucuronosyltransferase (UGT1A)

Function: transformation of small lipophilic molecules into water-soluble excretable metabolites

Diseases: sickle cell anemia, cancers, Gillbert's syndrome

Total	Event distribution				
Events	GA	OWM	NWM	PROS	DOG
23	0.00	0.09	**0.48**	0.17	0.26





Chemokine ligand genes (CCL)

Function: secreted proteins involved in immunoregulatory and inflammatory processes

Diseases: HIV susceptibility

Total	Event distribution				
Events	GA	OWM	NWM	PROS	DOG
9	**0.44**	0.33	0.00	0.11	0.11

(see 11 more CTAs below. . .)

References and acknowledgements

Dumas, L. et al. (2007). Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res*, 17(9):1266–1267.

NHGRI (2008). Approved sequencing targets. http://www.genome.gov/10002154.

Zhang, Y., Song, G., Vinař, T., Green, E. D., Siepel, A., and Miller, W. (2008). Reconstructing the evolutionary history of complex human gene clusters. In *RECOMB*, pages 29–49.

TV supported by the Packard fellowship awarded to AS.