Recent Duplications, Evolution and the Assembly of the Human Genome.

Evan Eichler
Case Western Reserve University

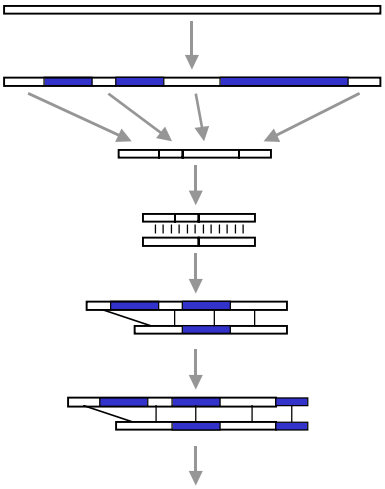
Human Genome Duplications

Question: How common are recent genome duplications in man? What are the molecular bases and consequences of these processes?

Hypotheses: Recent genomic duplications

- Show temporal and positional biases in genome structure
- Have created genes in man with new functions
- Are hotspots for rapid evolutionary change
- Are hotspots for recurrent chromosomal structural rearrangement.

METHOD



identify high copy repeats

splice out repeats

blast comparisons

reinsert repeats

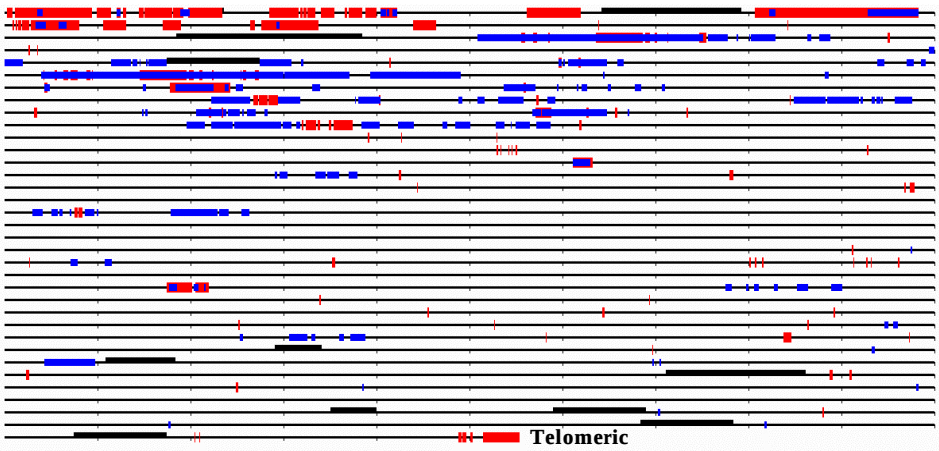
heuristic end trimming

global alignments

Analyze alignments
(>1 KB; >90% similar)

Duplication Structure of Chromosome 22q

Centromeric

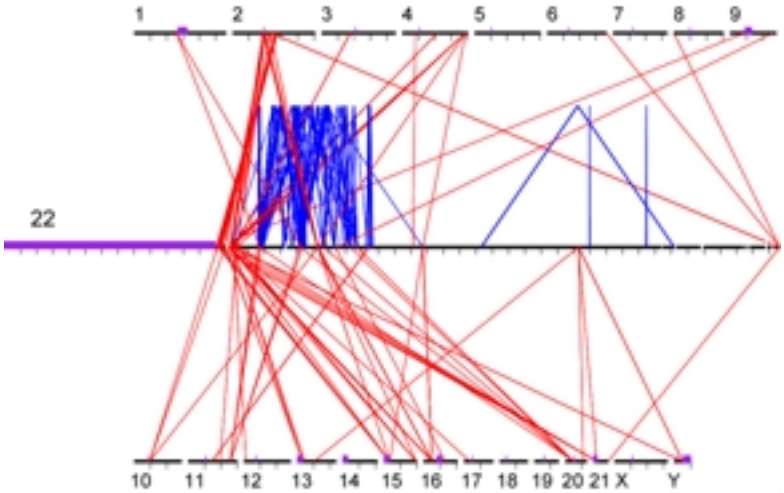


Telomeric

2011 Alignments

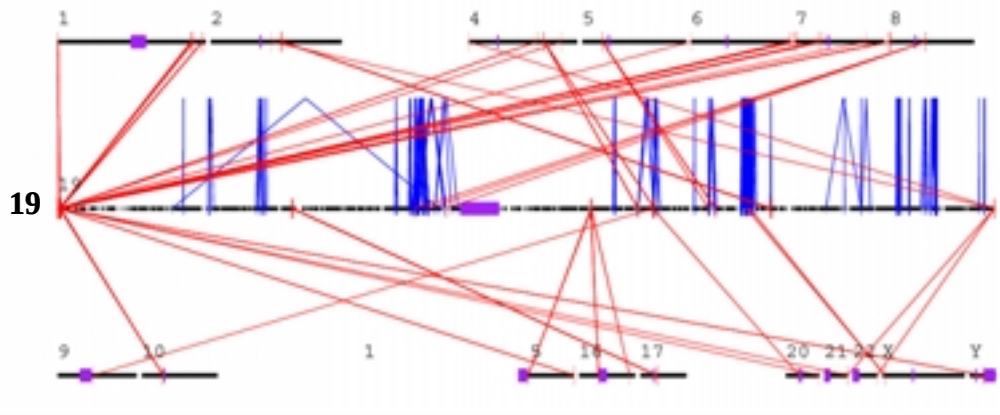
Total: 9.1% (3.04 Mb)
Inter: 3.9% (1.30 Mb)
Intra: 6.4% (2.19 Mb)

Pattern of Chromosome 22 Duplications



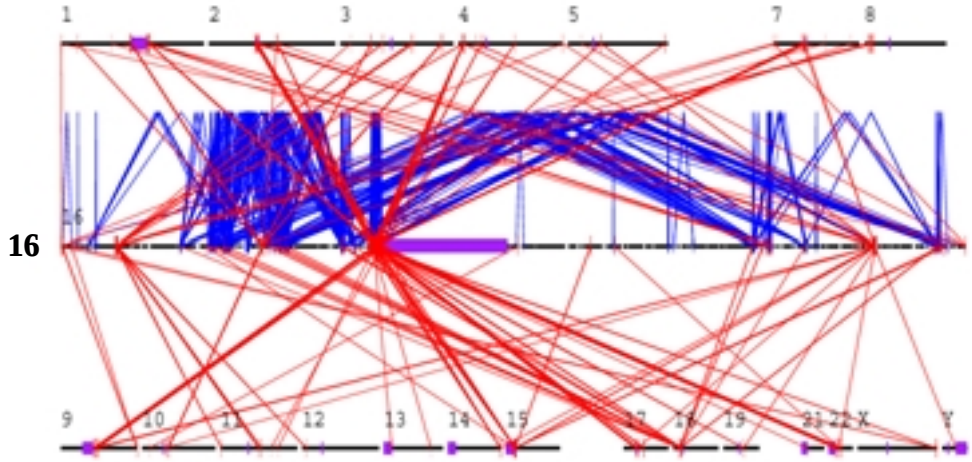
Total: 9.1%

Pattern of Chromosome 19 Duplications



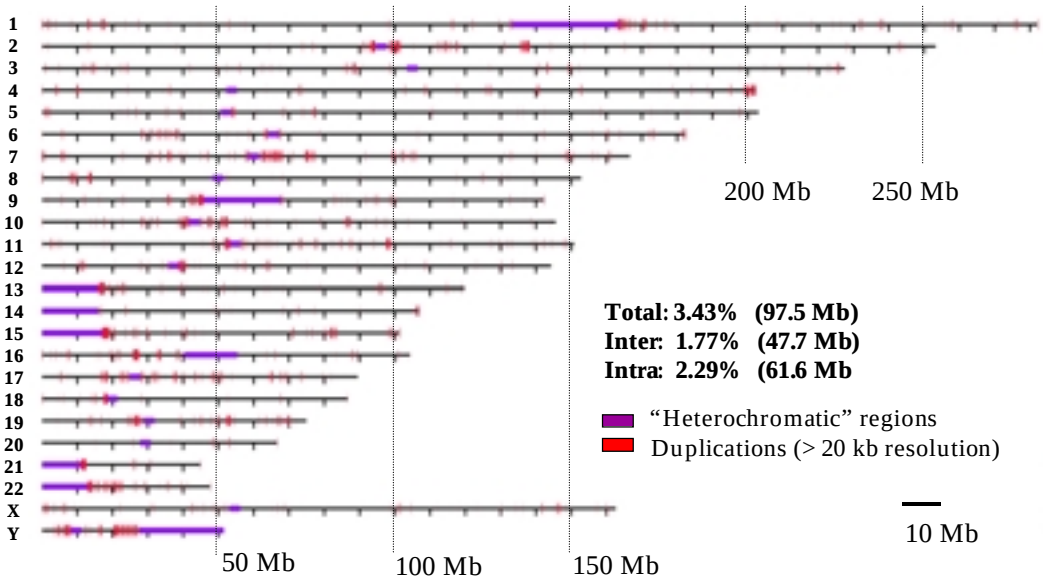
Total: 6.7%

Pattern of Chromosome 16 Duplications



Total: 10.3%

Segmental Duplications: (oo23, 90-98%, >5kb)



Total: 3.43% (97.5 Mb)
Inter: 1.77% (47.7 Mb)
Intra: 2.29% (61.6 Mb

"Heterochromatic" regions
Duplications (> 20 kb resolution)

25,169 Pairwise Alignments (>1kb)

Bailey *et al.* (2001), *Genome Res.*:11:1005-1017

Segmental Duplication: A Unique property of Human Genome Architecture

Duplicated Bases	FLY	WORM	Chrom 22	Total
> 1 KB	1.20%	4.25%	9.50%	3.64%
> 5 KB	0.37%	1.50%	7.90%	3.02%
>10 KB	0.08%	0.66%	6.40%	2.84%

•Genome structure of humans differs by an order of magnitude or more in terms of large duplications

•Important Biological and Practical Implications

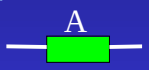
The Problem:

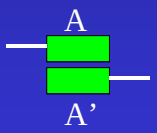
- Size >300 kb

2) Sequence Identity >95%

3) No Identifying Features.

Consequences:

1) Under-representation 

2) Misassembly 

Exceptional Targets of the Genome Project

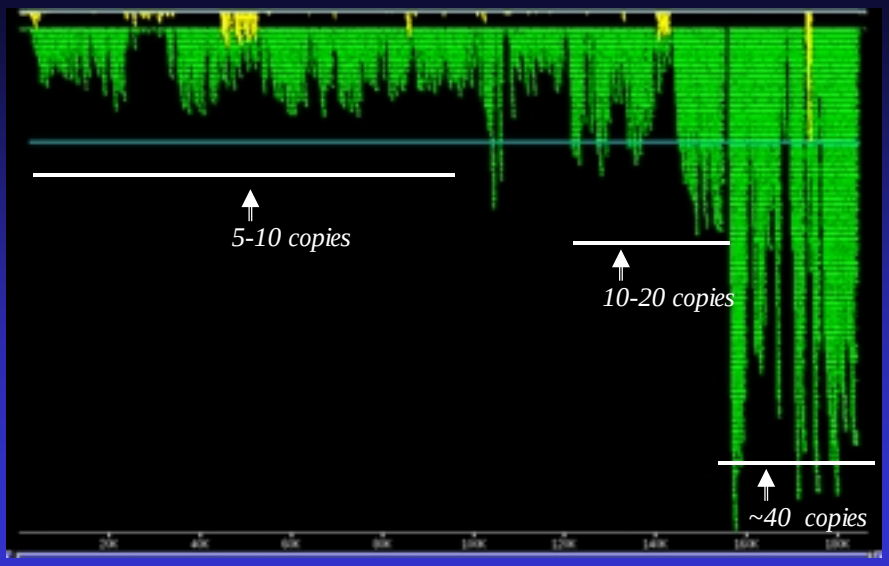
- Under-represented (~55% Coverage)
- Misassigned or Not assigned (~30%)
- Misassembled (particularly for segments >99%)
- SNPs vs Paralogous sequence variants (1.2-2X)
- Artifactual Duplications.

True Finishing of Human Genome will not simply entail topping off working draft sequence.

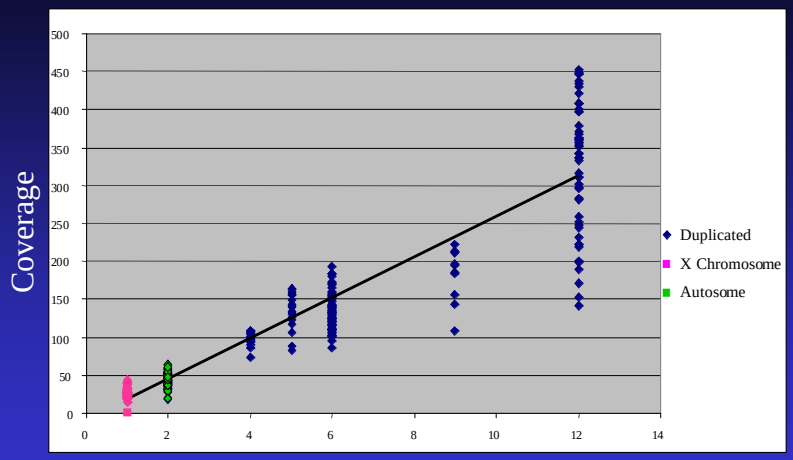
AC002038

Duplications >98% Sequence Identity.

Method: Celera Random Reads + Public Ordered Data



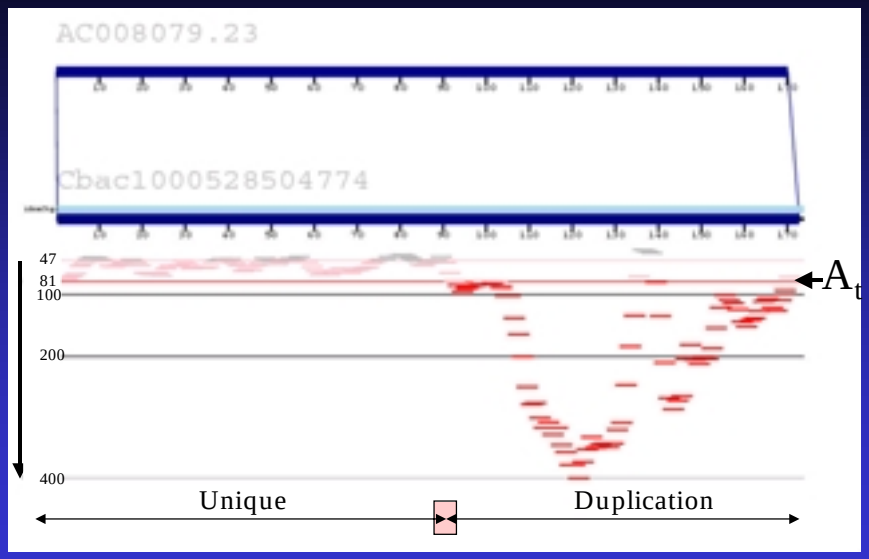
Depth of Coverage vs. Copy Number of Duplication



- dpt—5kb "unique"
- 16 unique acc.
- 23 paralogous acc.
- Ua=1641pts
- Ux=911 pts
- D= 4043 pts

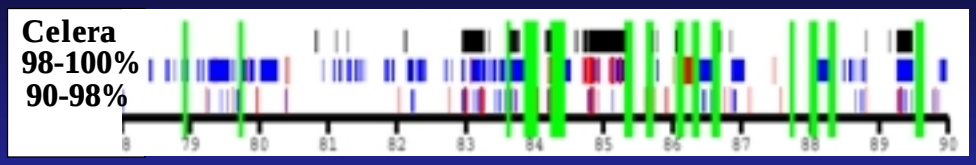
Copy # of Duplication

VCF Duplicon Transition (Unique to Repeat)



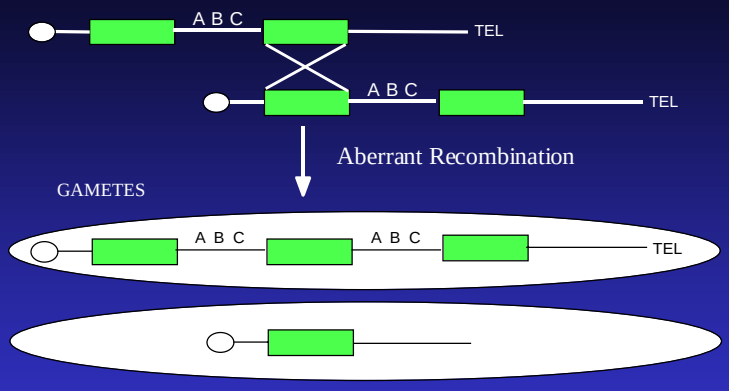
Strategy to Finish Duplicated Regions

- **Paralogy Detection**—*In silico* methods; use all HGP data.



- Experimentally characterize- target focal regions, monochromosomal source material to distinguish copies.
- Sequence—Identify clones that fill gaps.
- Reassemble—divide the genome into two segments highly homologous and possibly structurally polymorphic vs. unique regions of the genome.

Duplications and Disease



Human Disease

Triplosensitive, Haploinsufficient and Imprinted Genes

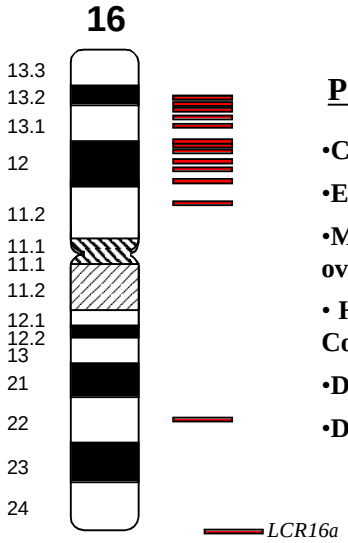
Duplication Mediated Recurrent Rearrangements

- Velocardiofacial / DiGeorge Syndome
- Angelman / Prader-Willi Syndrome
- Charcot-Marie Tooth Disease
- Smith-Magenis Syndrome
- Juvenile Nephronopthisis
- Panic and Phobic Disorders

Genome-wide screen for structural polymorphism associated with recent duplications

Duplications and Rapidly Evolving Genes

LCR16a (Low-copy repeats on chromosome 16)

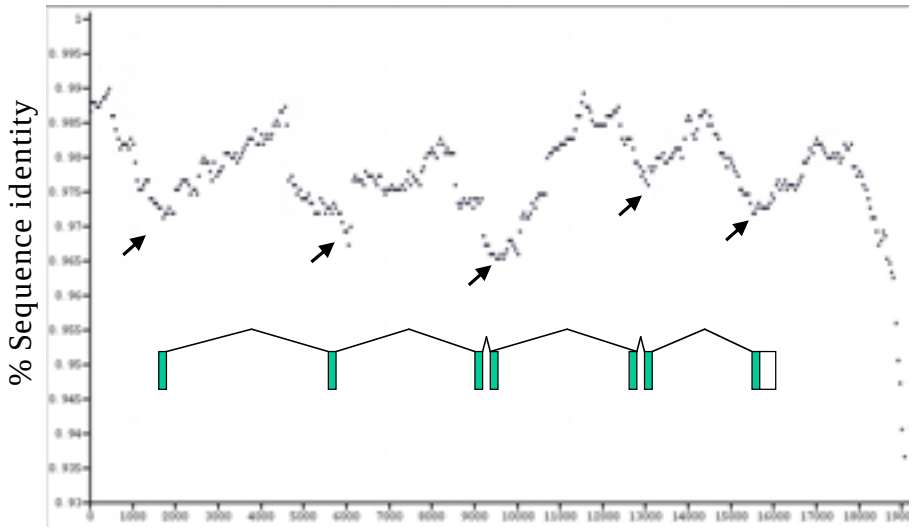


Properties:

- **Chromosome 16 specific**
- **Estimate copy number in Humans ~15-17 copies**
- **Majority are not tandem but are interspersed over >20 Mb of chromosome 16p.**
- **HGP–(15 distinct copies) BACs abut large gaps Copies have been misassigned, misassembled, etc**
- **Duplications are 19-20 kb in size**
- **Duplications are 97.5 to 99.5 % identical**

Alignment of Two Human LCR16a Copies

(window: 1500bp slide: 50 bp, ALIGNSLIDER)



Alignment of Putative Protein Sequences.

(Exon 2 – 4, LCR16a)

```
Human1 VINTLADHHH RGTDFGG--- --L------- --------LH VIIAFPTSYK VVITLWIVYL WVGNKIGLKD VITLRRHVET KVRAKIRKRK VTTKINHHDK INGKRKTARK
Human2 .....S.... .......--- --.------- --------.. .......... .......... .......... .......... .......... .......... .......... .
Human3 .......... .......--- --.------- --------.. .......... .......... ......V... .......... .......... .......... .......... .
Human4 ........R. .......--- --.------- --------.L I.TV.LR... FA.S.CTT.. ...I..V.E. IF..W.Q... ........M. ....V.R... ........KE
Human5 ........R. .E.....--- --.------- --------II I..V.LGR.. FT.LFCTI.. C..I....E. IC..WKQA.. ..Q.....M. ..K.V...Y. ........KE
Human6 ........R. .......--- --.------- --------.L I.TV.LR... FA.S.CTS.. C..I..V.E. IF..W.Q... ........M. ....V.R... ........KE
Human7 ...S..VYR. .E....VGVR DHPGQHGKTP SPQKLDNLII I..G.LRR.T FN.LFCTSC. C......... .......... .......... ......R... ..........

Chimp2 ......VY.. .E...C.GVR DHPGQHGKTP SPQKLDNLII I..G.LRP.T FT.LFCTN.. C......... ...PW....R .F....H... .......... ..........

Chimp1 .......LR. .......--- --F------- --------.L ISTV.LR... FA.S.CTS.. S..I..V.EA IF..W.Q.K. ......H.M. ....V.R... .......TKE
Chimp3 ........R. .......--- --.------- --------.R I....LR... ..V...T... .......... I...W..... ......H.M. .......... ........K.
Chimp4 .......LR. .......--- --F------- --------.L I.TV.LR... FA.S.RTS.. S......R.. ........KR ......H.M. .......... ........K.
Chimp5 ........R. .......--- --.------- --------.G I..E.LR... ......T... .......R.. ........K. ......H.M. ....V..... ........K.

Gibbon1..HS.P.DF. P....C.--- --.------- --------IV I..V.LGIST LA.F..KTS. C..I.F.RE. RF.SW..M.A ....EV.--- ..G.V.S.Y. ...Q....EE
Giboon2...SPP..F. P....C.--- --.------- --------IV I..V.LGIST LA.F..KTS. C......R.. ....W...K. ..TP....M. .......... .S......K.
Gibbon3...SPP..F. P....C.--- --.------- --------IV I..V.LGIST LA.F..KTS. C..I....E. LF.SQ..M.A ....EVH--- ..KNV.S.Y. ...H....K.
Patas  ...D.P..FQ ..P..F.--- --.------- --------RA I..V.LGIST LG.F..KTSF G..I..A.E. LF.S.SFM.A RA..EVH--- ..R.V.S.YQ ...Q...TE.
Baboon ...S.P..FQ P.P..F.--- --.------- --------IA I..V.LGIST LG.F..KASF G..I....E. LF.SQ..M.A ....EVH--- ..RNV.S.Y. ...H....K.
Colobus...S.P..FQ P.P..F.--- --.------- --------IA I..V.LGIST LG.F..KTSF G..I...RE. LF.SW.YM.A ....VVH--- ....V.S.YQ .H.Q...TEE
```

- 20-25% amino acid change between Human / Chimp (2-3% nt)
- 40-45% amino acid change between Human / OW (7-8% nt)

Summary.

1) Human Genome two types of recent paralogy:
   a) Interchromosomal paralogy
   b) Chromosome-specific paralogy
2) ~5% of Human Genome
3) Problematic for sequence and assembly.
4) Paralogy-detection method: Combined public and private to identify duplicated regions
5) Constructed a Human Paralogy Database of duplicated Segs.
6) Important for evolution of new gene function and disease.

Acknowledgements

Amy Yavor Julie Horvath Meghan Smith Devin Locke
Jeff Bailey Christie O'Keefe Matt Johnson Uli Neuss

HGP
David Haussler
Richard Gibbs
Shaying Zhao
John Mcpherson
Anne Olsen
Laurie Gordon
Norman Doggett
Pieter DeJong

Celera
Peter Li
Mark Adams
Zhiping Gu
Gene Myers
Knut Reinert