Analysis of large deletions in human-chimp genomic alignments

Erika Kvikstad

BioInformatics I

December 14, 2004

Outline

- Mutations, mutations, mutations…
- Project overview
  - Strategy: finding, classifying indels
  - Anaylsis
- Future Goals

Why study mutations?

- sources for genetic disease
  - 68% nucleotide substitutions
  - 23% small indels

from Human Gene Mutation Database

- genetic variation in natural populations
- stuff of molecular evolution

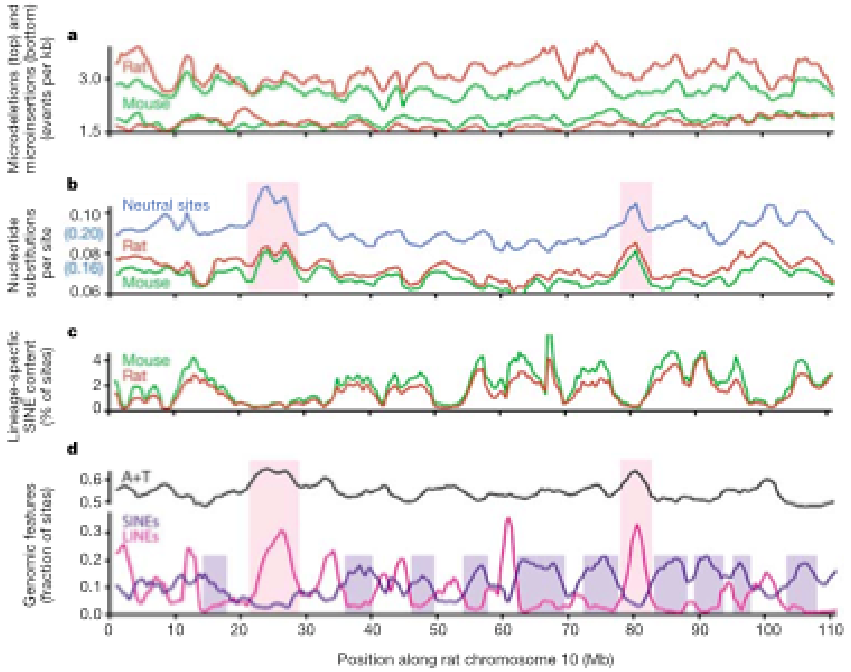
Variability in mutation rate

Factors

- GC content
- Recombination

Scales

- DNA sequence context
- within chromosomes
- between chromosomes



Rat Genome Sequencing Project Consortium*

NATURE| VOL 428| 1 APRIL 2004 |www.nature.com/nature

Why indels in particular?

- Indels responsible for more unmatched nucleotides between closely related populations:
  - Drosophila
  - Arabidopsis
  - Sea urchin
  - Primates
  - E. coli O157:H7

(Britten, PNAS, 2003, Vol 100)

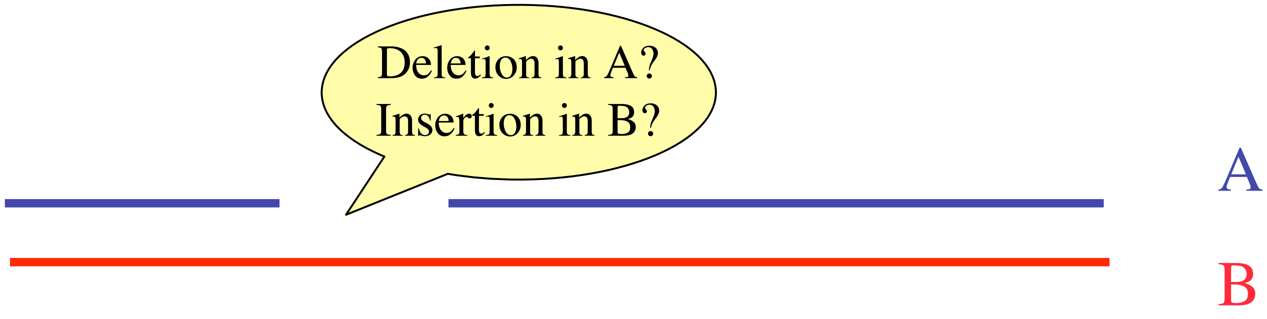
Human vs Chimp

- Use human as reference
  - Most complete sequence from large-scale genome project (IHGSC, Nature, 2004, Vol.431)
    - assembly issues minimal
    - Alignment gaps infer true deletion event
- Compare to chimpanzee sequence
  - Nearest primate relative

    Divergence from human established from sequence
      - substitutions 1.2 to 1.4%
      - indel rate 5% (Britten, PNAS, 2002, Vol99)

Model to Identify Indels

Gaps in alignment of two sequences



- How to determine history of this event?

Gaps in an Alignment of 2 sequences

Alignment

Assembly - missing sequence data

Indels: inferring ancestral state

Insertion vs Deletion

- Methods
  - Use phylogenetic history: outgroup
  - Use model: retrotransposition
    - 80%> repeat content of indel

      and repeat <5% divergence from consensus to infer insertions
      - Used in previous literature (e.g. Liu, Genome Research, 2003, Vol.13)

Transposition - Mechanisms

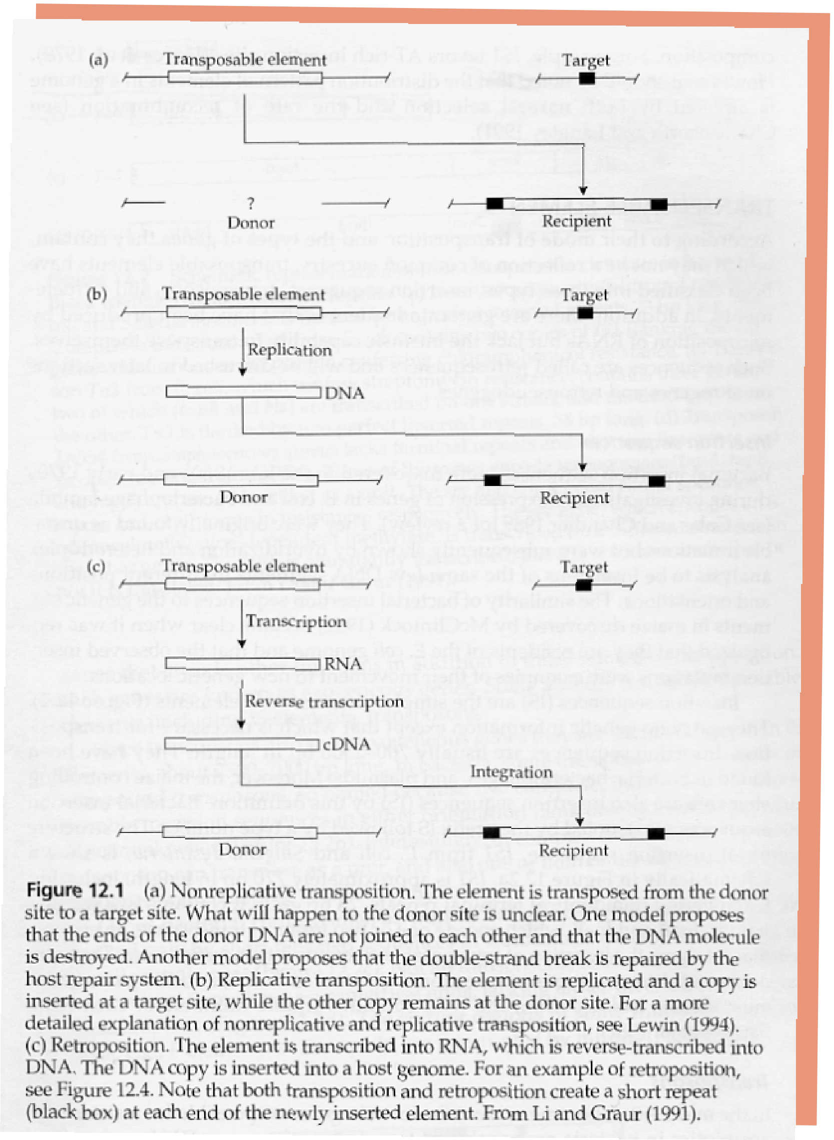


Figure 12.1 (a) Nonreplicative transposition. The element is transposed from the donor site to a target site. What will happen to the donor site is unclear. One model proposes that the ends of the donor DNA are not joined to each other and that the DNA molecule is destroyed. Another model proposes that the double-strand break is repaired by the host repair system. (b) Replicative transposition. The element is replicated and a copy is inserted at a target site, while the other copy remains at the donor site. For a more detailed explanation of nonreplicative and replicative transposition, see Lewin (1994). (c) Retroposition. The element is transcribed into RNA, which is reverse-transcribed into DNA. The DNA copy is inserted into a host genome. For an example of retroposition, see Figure 12.4. Note that both transposition and retroposition create a short repeat (black box) at each end of the newly inserted element. From Li and Graur (1991).

Interspersed Repeats

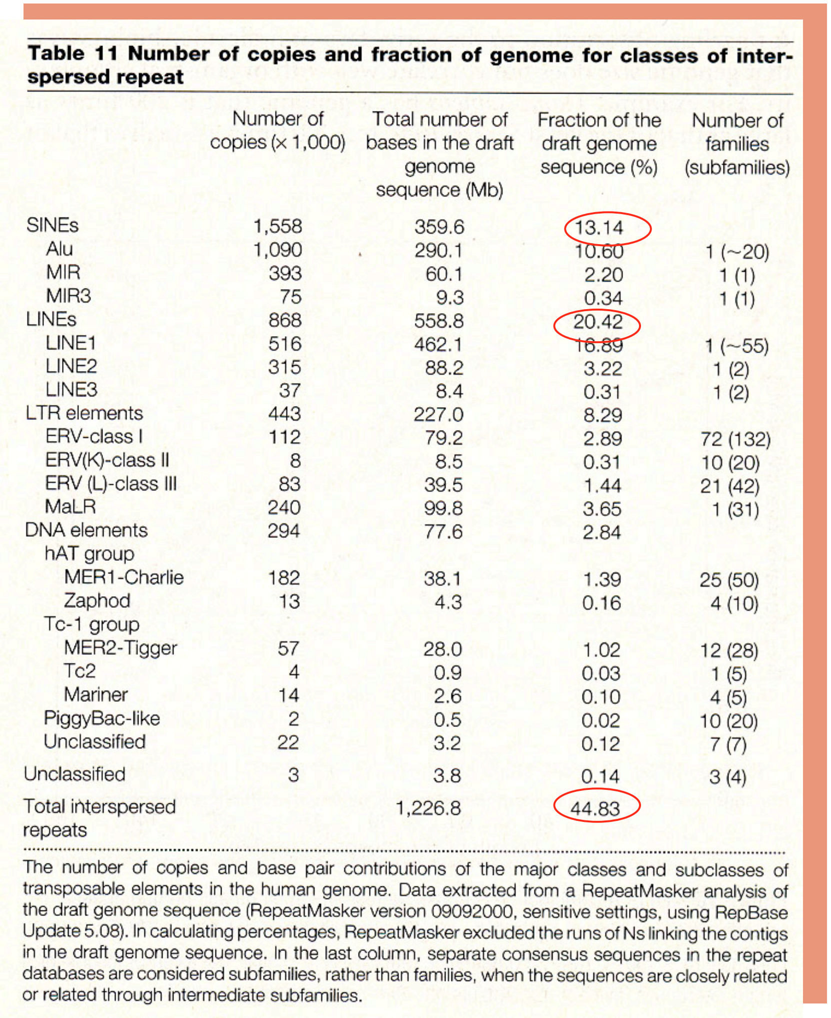


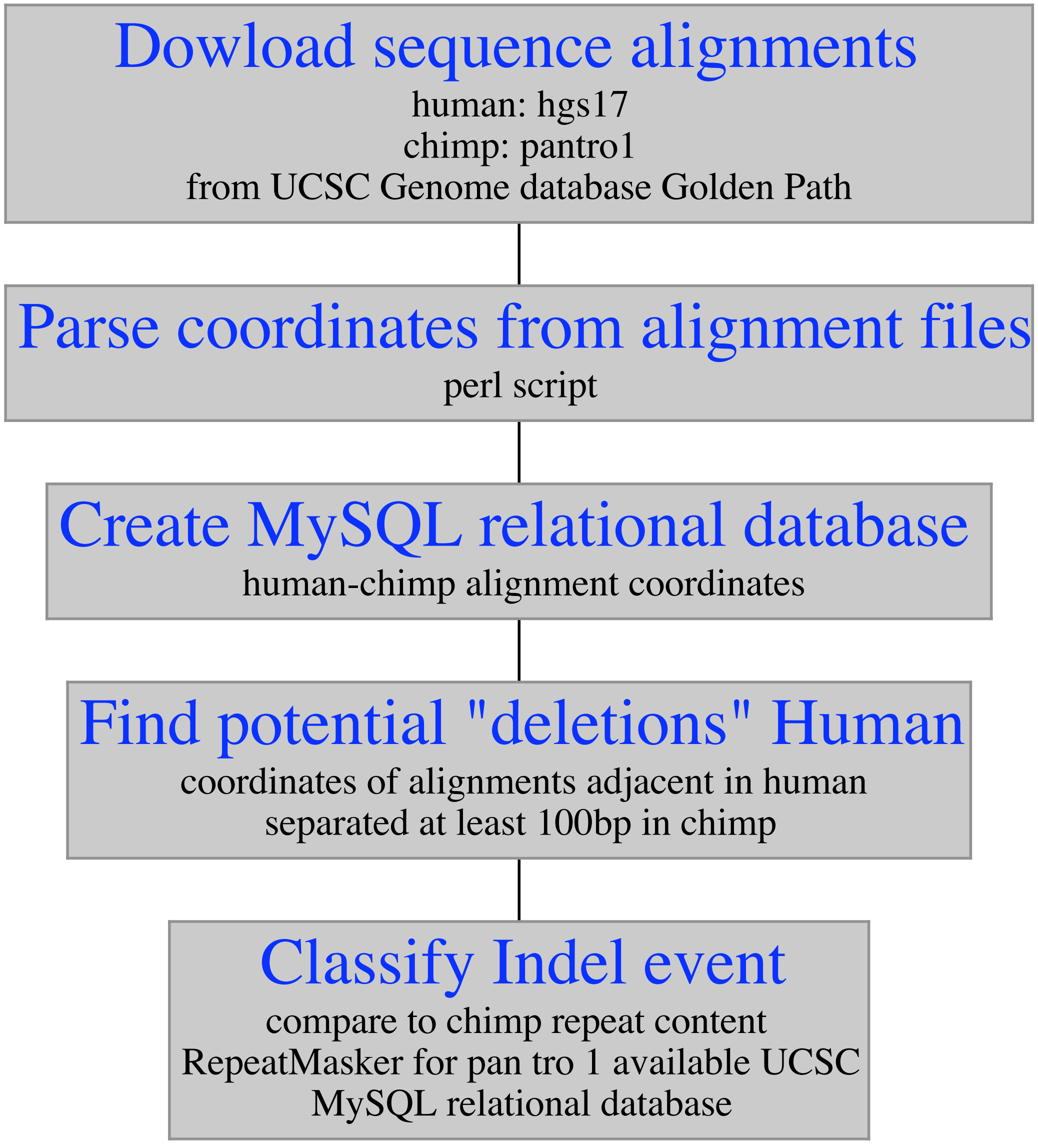
Table 11 Number of copies and fraction of genome for classes of interspersed repeat

	Number of copies (× 1,000)	Total number of bases in the draft genome sequence (Mb)	Fraction of the draft genome sequence (%)	Number of families (subfamilies)
SINEs	1,558	359.6	13.14	
Alu	1,090	290.1	10.60	1 (~20)
MIR	393	60.1	2.20	1 (1)
MIR3	75	9.3	0.34	1 (1)
LINEs	868	558.8	20.42	
LINE1	516	462.1	16.89	1 (~55)
LINE2	315	88.2	3.22	1 (2)
LINE3	37	8.4	0.31	1 (2)
LTR elements	443	227.0	8.29	
ERV-class I	112	79.2	2.89	72 (132)
ERV(K)-class II	8	8.5	0.31	10 (20)
ERV (L)-class III	83	39.5	1.44	21 (42)
MaLR	240	99.8	3.65	1 (31)
DNA elements	294	77.6	2.84	
hAT group				
MER1-Charlie	182	38.1	1.39	25 (50)
Zaphod	13	4.3	0.16	4 (10)
Tc-1 group				
MER2-Tigger	57	28.0	1.02	12 (28)
Tc2	4	0.9	0.03	1 (5)
Mariner	14	2.6	0.10	4 (5)
PiggyBac-like	2	0.5	0.02	10 (20)
Unclassified	22	3.2	0.12	7 (7)
Unclassified	3	3.8	0.14	3 (4)
Total interspersed repeats		1,226.8	44.83	

The number of copies and base pair contributions of the major classes and subclasses of transposable elements in the human genome. Data extracted from a RepeatMasker analysis of the draft genome sequence (RepeatMasker version 09092000, sensitive settings, using RepBase Update 5.08). In calculating percentages, RepeatMasker excluded the runs of Ns linking the contigs in the draft genome sequence. In the last column, separate consensus sequences in the repeat databases are considered subfamilies, rather than families, when the sequences are closely related or related through intermediate subfamilies.

IHGSC, Nature, 2001

Human Deletions Strategy



Dowload sequence alignments

human: hgs17
chimp: pantro1
from UCSC Genome database Golden Path

Parse coordinates from alignment files

perl script

Create MySQL relational database

human-chimp alignment coordinates

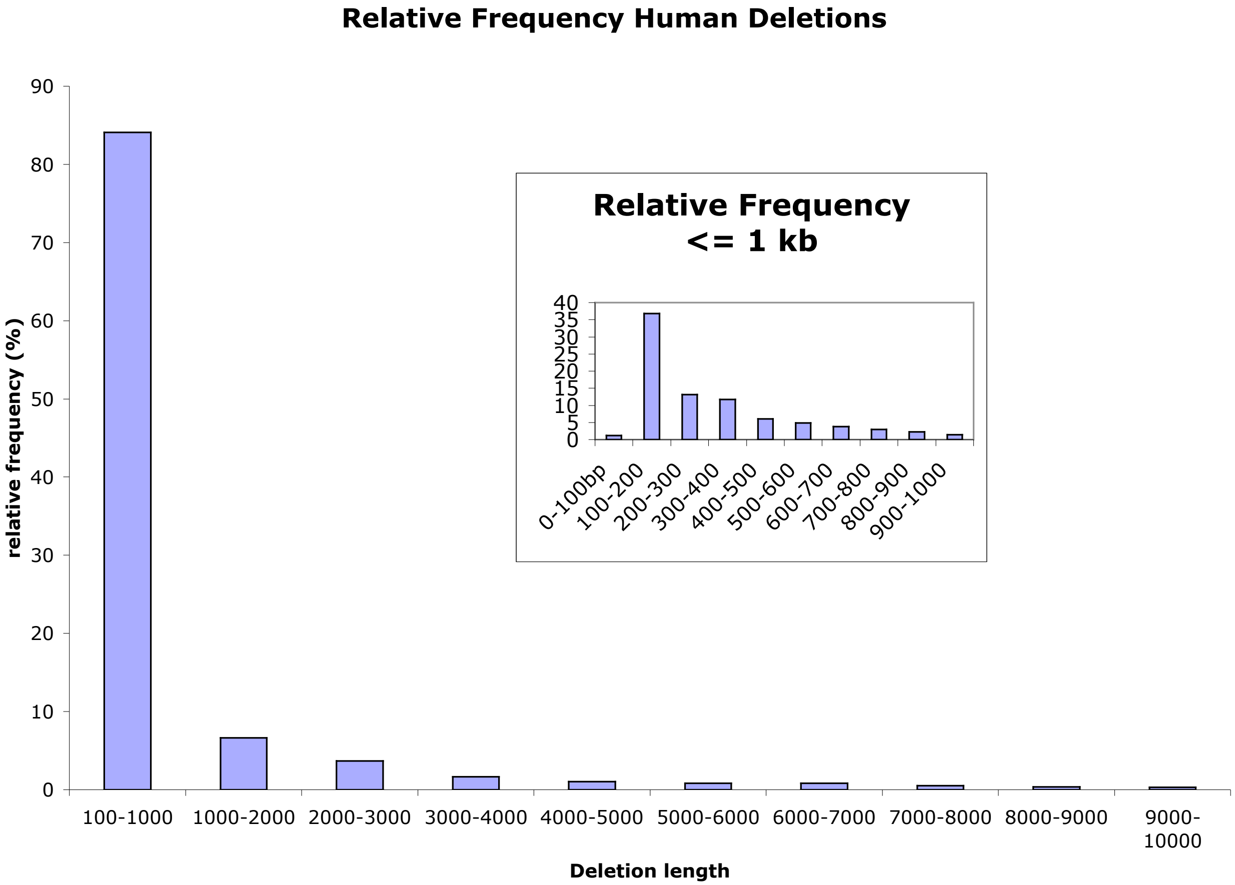
Find potential "deletions" Human

coordinates of alignments adjacent in human
separated at least 100bp in chimp

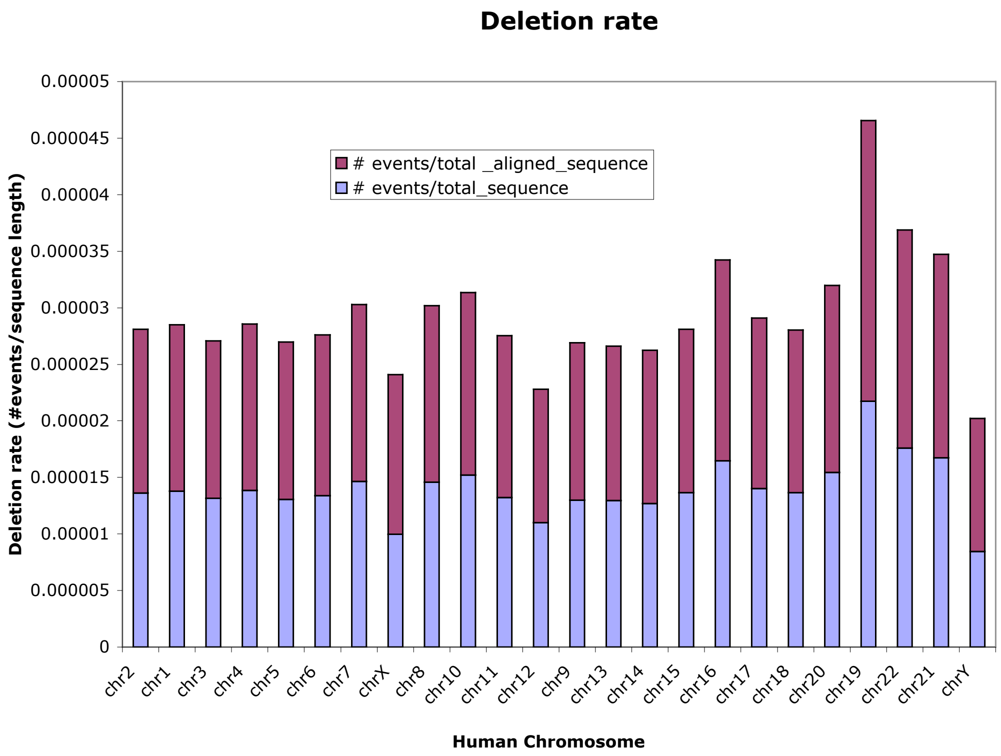
Classify Indel event

compare to chimp repeat content
RepeatMasker for pan tro 1 available UCSC
MySQL relational database

Results


Relative Frequency Human Deletions
relative frequency (%)
90
80
70
60
50
40
30
20
10
0
100-1000
1000-2000
2000-3000
3000-4000
4000-5000
5000-6000
6000-7000
7000-8000
8000-9000
9000-10000
Deletion length
Relative Frequency
<= 1 kb
40
35
30
25
20
15
10
5
0
0-100bp
100-200
200-300
300-400
400-500
500-600
600-700
700-800
800-900
900-1000

…cont.



Total sequence length data from (IHGSC, Nature, 2004, Vol.431)

- •Castresana (2002) reported significantly higher Ks for hum chr 19 than any other hum chr
  - —based on human/mouse orthologous gene pairs

- •Chr 19 unusual
  - High GC content
  - High gene density
  - High expression levels

(Castresana, Nucl Acid Res, 2002, Vol 30)

Discussion

Compare to chimp chr22 paper results:

(analysis of indels 300-5,000bp range)

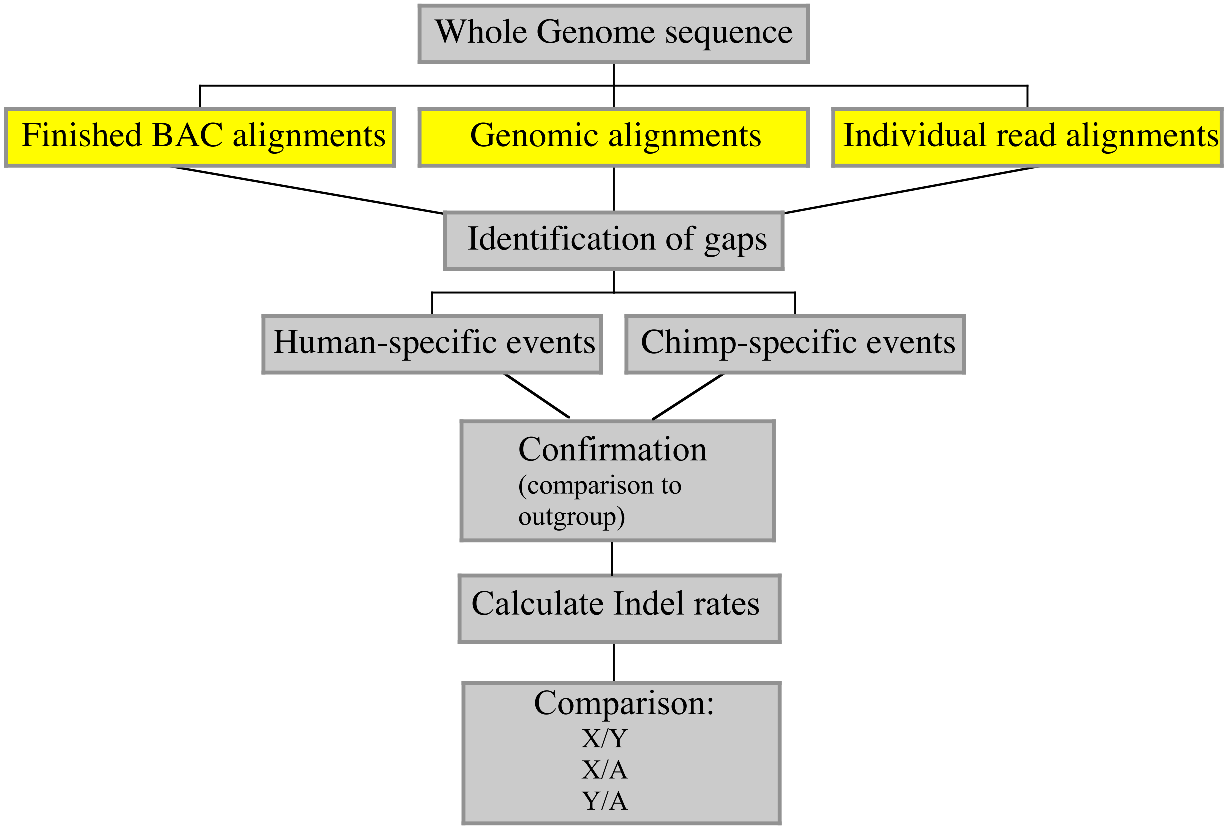
	Int'lChimpChr22 Consort.	**Deletion analysis**
PTR22 Inserted	25kb	7kb
HAS 21 deleted	39kb	268kb
Indels total	567	270

(ICC22C, Nature, 2004 Vol429)

Future

- Definition of deletion
  - Allow for“wobble” alignment ends
    - +1, +5, +20 bp threshold for adjacent human sequence
- Size distributions
  - Different T thresholds
    - e.g. 10, 50, 100 kb etc. threshold for lower limit indel
- Filter data
  - Errors due to low coverage chimp sequence
    - e.g. RepeatMasker chimp
    - e.g. Bias certain chromosomes sequenced to better completion than others assembly - repetitive regions require finishing strategies
  - Compare to other alignment sources

Indel strategy



Whole Genome sequence

Finished BAC alignments

Genomic alignments

Individual read alignments

Identification of gaps

Human-specific events

Chimp-specific events

Confirmation
(comparison to
outgroup)

Calculate Indel rates

Comparison:
X/Y
X/A
Y/A

Future

- Definition of deletion
  - Allow for“wobble” alignment ends
    - +1, +5, +20 bp threshold for adjacent human sequence
- Size distributions
  - –Different T thresholds
    - •e.g. 10, 50, 100 kb etc. threshold for lower limit indel
- Filter data
  - –Errors due to low coverage chimp sequence
    - •e.g. RepeatMasker chimp
    - •e.g. Bias certain chromosomes sequenced to better completion than others assembly - repetitive regions require finishing strategies
  - –Compare to other alignment sources
- Underlying mechanisms
  - –Replication-driven vs Recombination-driven
    - •Different sizes of indels
    - •Chromosomal bias
- Other mammals
  - –Complete genome sequence available for mouse and rat

Acknowledgements

People:

Kateryna Makova

Webb Miller

Alignments and tools:

UCSC Genome Browser (http://genome.ucsc.edu)

Chimp Genome Sequencing Consortium

International Human Genome Sequencing Consortium

Sequence Alignments

The scoring matrix used for blastz was:

A	C	G	T
100	-300	-150	-300
-300	100	-300	-150
-150	-300	100	-300
-300	-150	-300	100

with a gap open penalty of 400 and a gap extension penalty of 30.

The alignments were done with blastz, which is available from Webb Miller's group at PSU. Each chromosome was divided into 10000000 base chunks with 10000 bases of overlap.

The axtNet alignments were processed with chainNet, netSyntenic, and netClass from Jim Kent at UCSC.

(http://genome.ucsc.edu)

Sequence Assemblies:

human/chimp alignments made using the May 2004 human assembly (hg17) vs. the Nov 2003 chimp assembly (panTro1) produced by the Chimp Genome Sequencing Consortium.

Repeats

RepeatMasker of Nov 2003 chimp assembly performed at the -s sensitive setting.

(http://genome.ucsc.edu)

Summary Statistics

hum_chrom	nt deleted	# deletions	# chimp insertions
chr2	2,282,021	3,237	160
chr1	2,431,061	3,073	144
chr4	1,973,522	2,592	108
chr3	1,845,042	2,559	148
chr5	1,600,512	2,319	69
chr7	1,706,268	2,265	88
chr6	1,567,469	2,242	132
chr8	1,556,725	2,081	98
chr10	1,563,384	2,000	82
chr11	1,390,982	1,735	82
chr9	1,056,252	1,528	69
chrX	1,166,142	1,498	29
chr12	1,001,185	1,433	64
chr16	946,623	1,300	37
chr13	1,014,162	1,238	58
chr19	992,812	1,212	22
chr14	815,757	1,120	63
chr15	918,207	1,111	31
chr17	895,303	1,091	17
chr18	663,871	1,019	53
chr20	625,877	919	31
chr22	407,996	612	17
chr21	321,388	572	17
chrY	326,637	210	6

Total <= 10 kb

n	sum(bp)	min	max	mean
38,966	29,069,198	98	9,980	746.0144

<= 300b

n = 19,928 => 51%