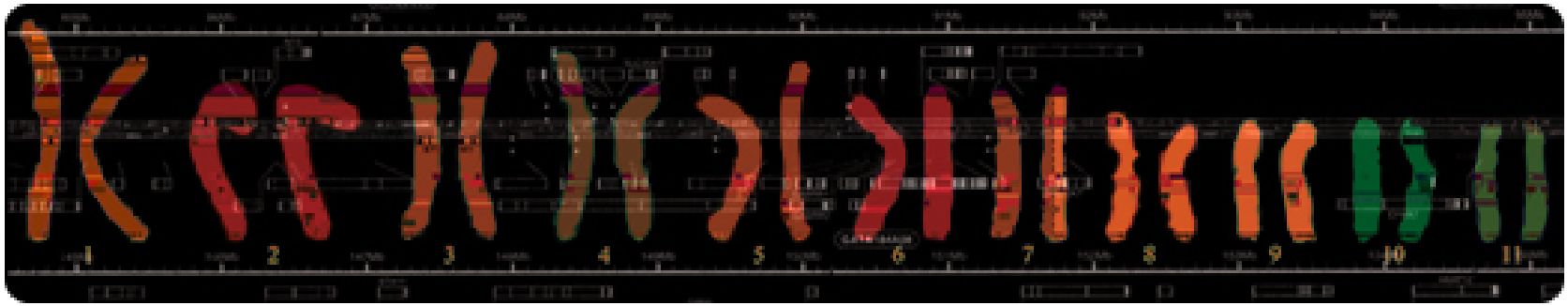


The Diploid Genome Sequence of an Individual Human

Maido Remm

Journal Club
12.02.2008

Outline

- Background (history, assembling strategies)
- Who was sequenced in previous projects
- Genome variations in J. Graig Venter genome
- Experimental validation of variations
- Tools for genome comparisons
- Conclusions

History of sequencing human genomes

26th June 2000 – Two different groups of scientists (Human Genome Project initiative and Celera Inc.) present their draft version of human genome.

15th February 2001 - Nature publishes draft human genome by the HGP group

16th February 2001 - Science publishes Celera's draft human genome

14th April 2003 - HGP consortium announces that sequencing of the human genome is finished

17th February 2004 – Venter group publishes two genome assembly comparisons in PNAS

21th October 2004 – An article in Nature publishes finished version of the human genome

4th September 2007 - The Diploid Genome Sequence of an Individual Human

Sequencing of the human genome 1988-2004

Two competing groups, two different approaches:

Human Genome Project – public (E.Lander)

Celera – commercial (J Craig Venter)

Terminology

Cloning tools:

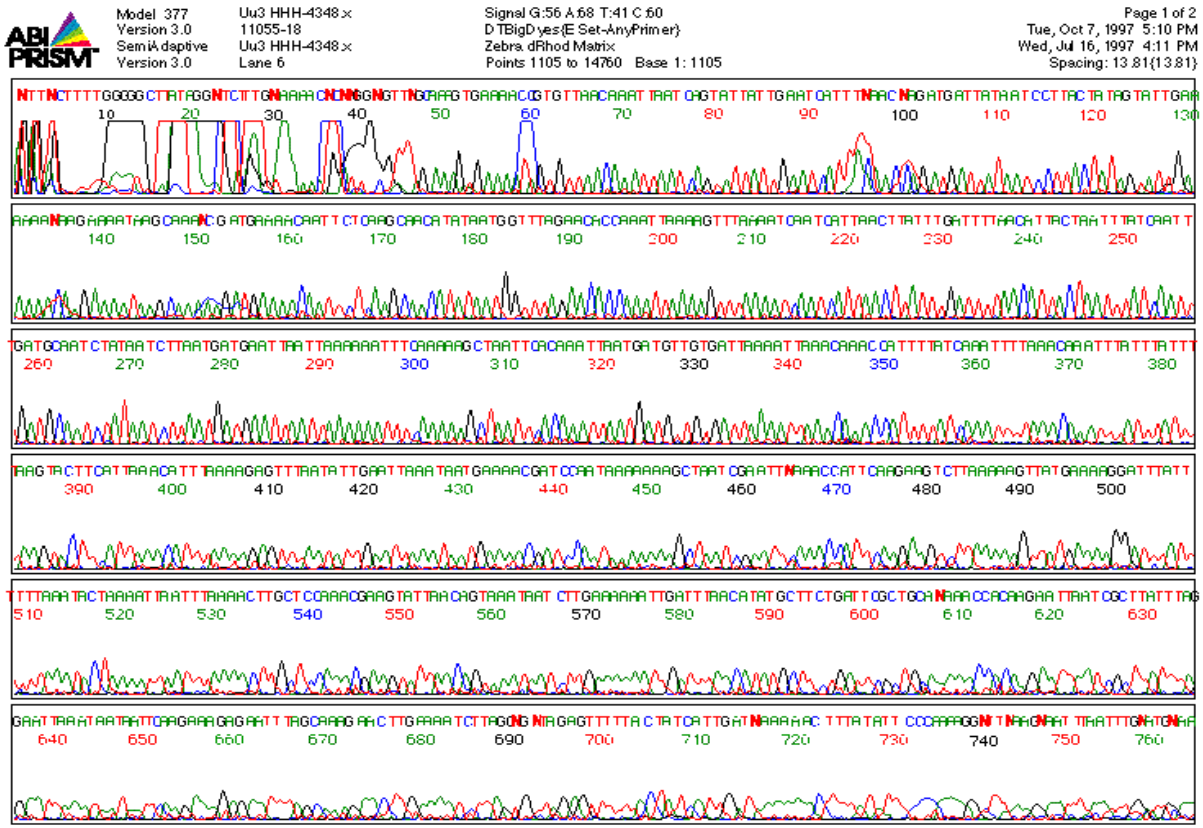
YAC (1Mb foreign DNA)
BAC clones (100 kb foreign DNA)
plasmid clones (1-10 kb foreign DNA)

Assembly steps:

read (ca 500 – 1000 bp sequence from clone ends)
contig (assembled from overlapping reads)
scaffold (assembled using the read pairs)
assembly (genome sequence assembled using the STS markers)

size

Sequence assembly: example of sequence trace (raw read)

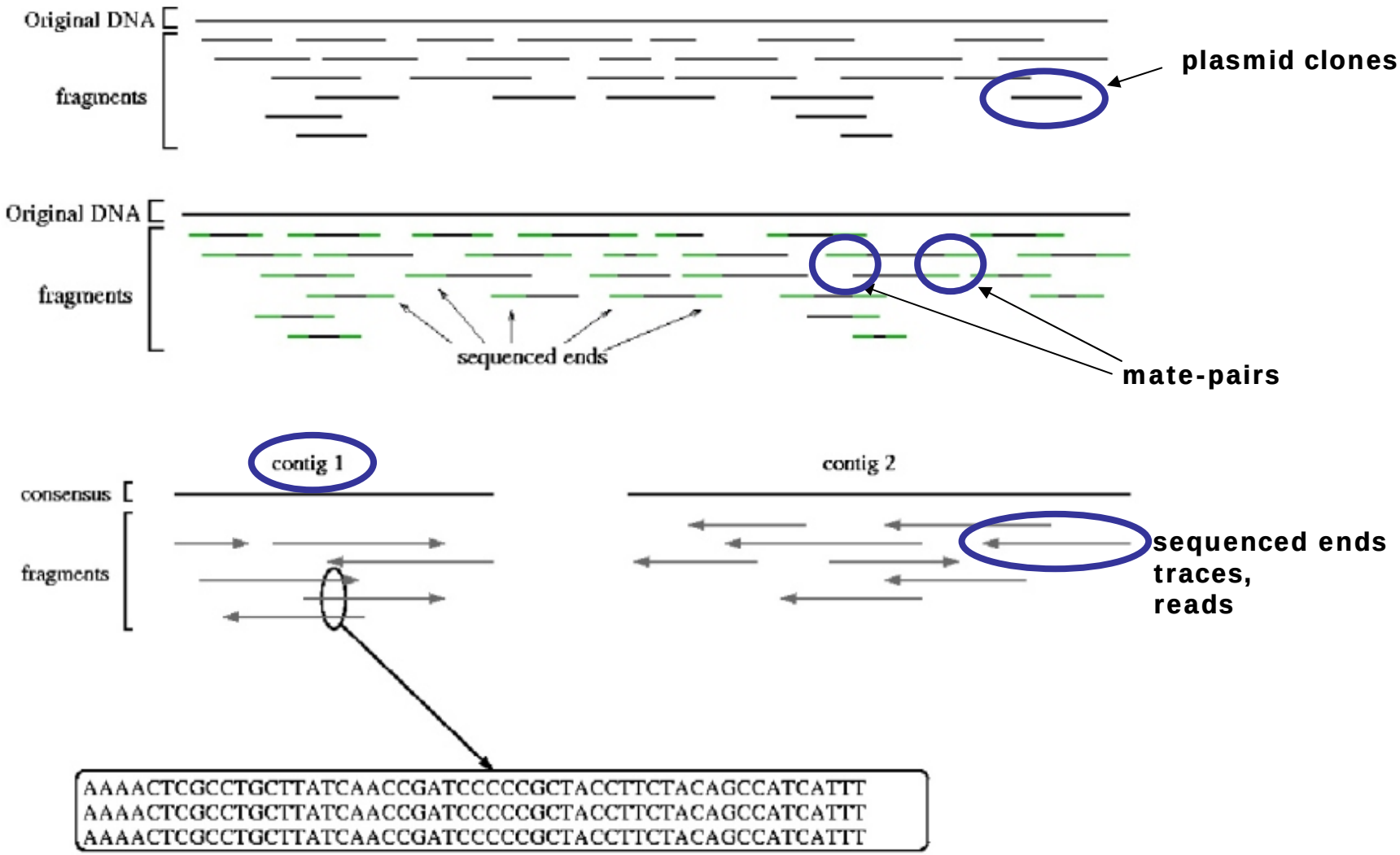


NB! Trace data is full of errors, particularly at ends.
Potentially useful source for discovery of new genomic features (inversions, viral sequences).

Trace database is available at http://www.ncbi.nlm.nih.gov/Traces/

Assembly of genomes

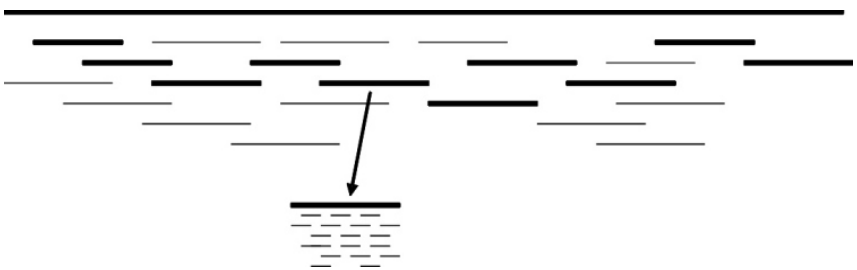
Assembly of bacterial genomes and BACs (100kb - 10 Mb sequences)

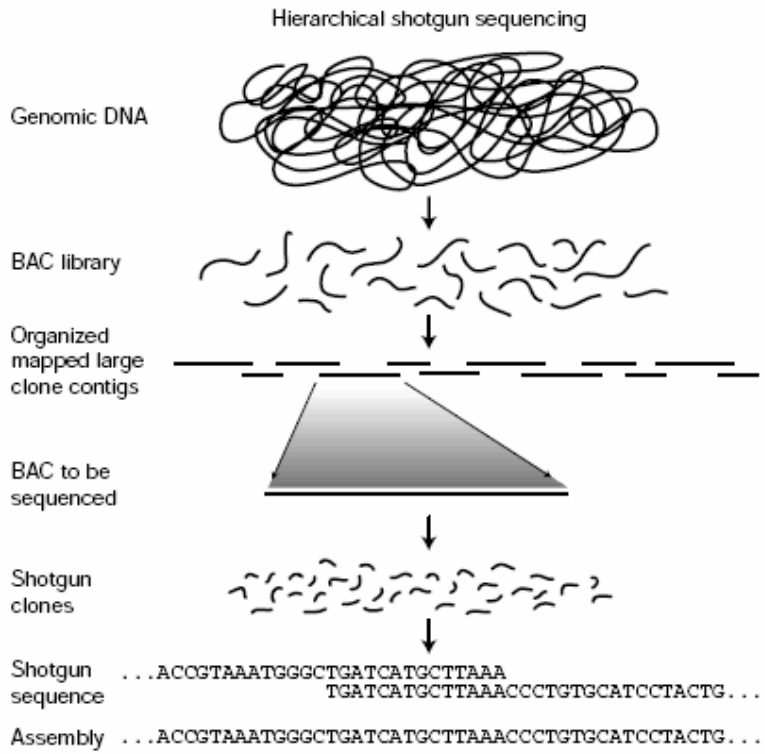


Assembly of eukaryotic genomes

Assembly of the human genome in Human Genome project:

BAC library was created, locations of BAC clones were determined by time-consuming mapping techniques.
Clones of BAC library were hierarchically cloned into plasmid libraries and sequenced in different sequencing centers.





Assembly of eukaryotic genomes by WGSA

(invented in Celera)

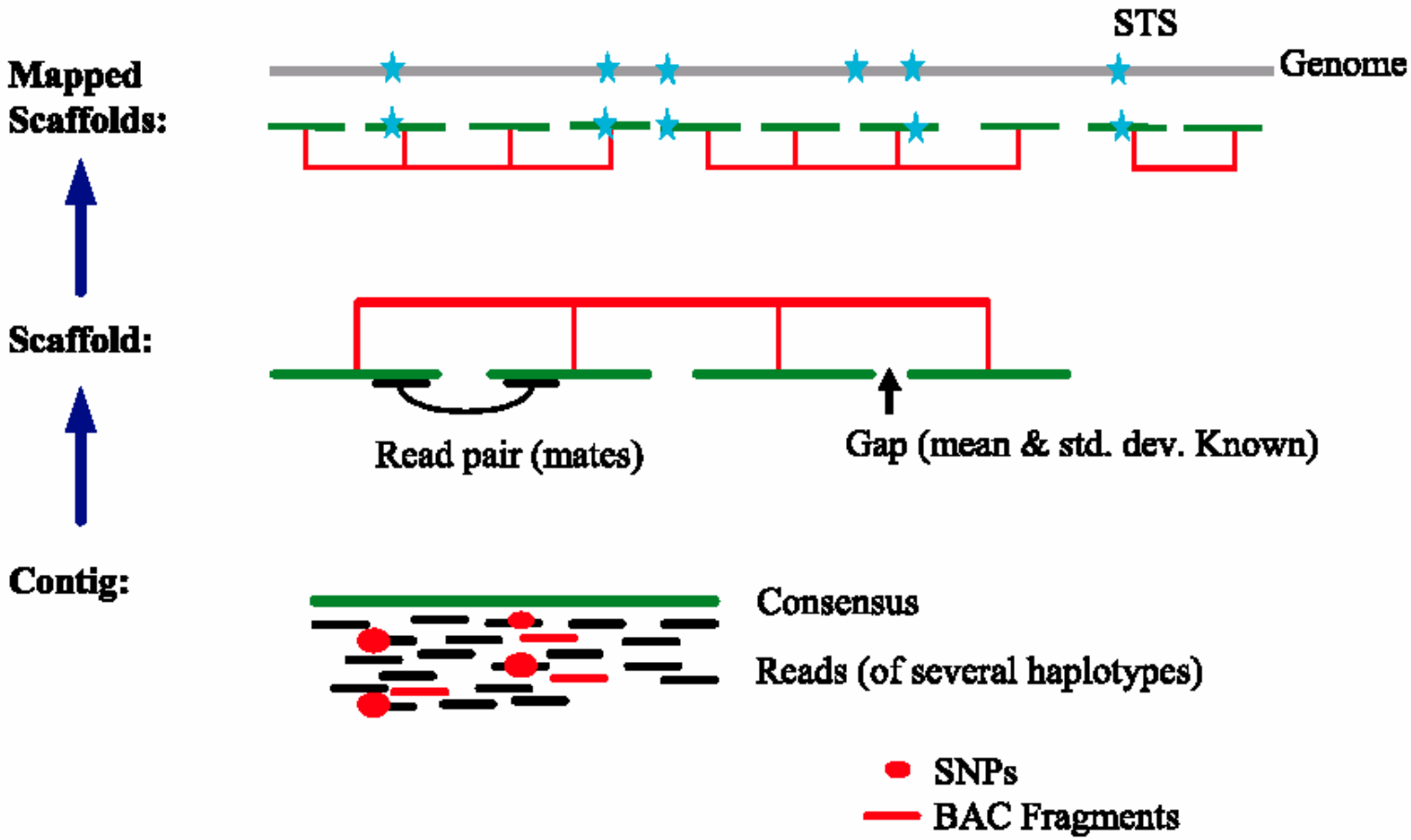


Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

Who was sequenced?

In Human Genome Project:

Many anonymous DNA donors, volunteers were taken on first-come-first served basis.

From each donor's DNA one BAC library (kloonide raamatukogu) was created.

BAC clones were mapped to the genome

Most of the genome was covered by using 8 librarys.

BAC libraries were divided by participating centers: 20 sequencing centers or universities from 6 different countries.

NB! Most (74%) of public DNA sequence (NCBI_36) is sequenced from single BAC library (one individual) RPCI-11. It is NOT a consensus sequence of known human individuals.

Who was sequenced?

In Human Genome Project:

Table 1 Key large-insert genome-wide libraries

Library name*	GenBank abbreviation	Vector type	Source DNA	Library segment or plate numbers	Enzyme digest	Average insert size (kb)	Total number of clones in library	Number of fingerprinted clones†	BAC-end sequence (ends/clones/ clones with both ends sequenced)‡	Number of clones in genome layout§	Sequenced clones used in construction of the draft genome sequence		
											Number‖	Total bases (Mb)¶	Fraction of total from library
Caltech B	CTB	BAC	987SK cells	All	*Hin*dIII	120	74,496	16	2/1/1	528	518	66.7	0.016
Caltech C	CTC	BAC	Human sperm	All	*Hin*dIII	125	263,040	144	21,956/ 14,445/ 7,255	621	606	88.4	0.021
Caltech D1 (CITB-H1)	CTD	BAC	Human sperm	All	*Hin*dIII	129	162,432	49,833	403,589/ 226,068/ 156,631	1,381	1,367	185.6	0.043
Caltech D2 (CITB-E1)		BAC	Human sperm	All									
				2,501–2,565	*Eco*RI	202	24,960						
				2,566–2,671	*Eco*RI	182	46,326						
				3,000–3,253	*Eco*RI	142	97,536						
RPCI-1	RP1	PAC	Male, blood	All	*Mbo*I	110	115,200	3,388		1,070	1,053	117.7	0.028
RPCI-3	RP3	PAC	Male, blood	All	*Mbo*I	115	75,513			644	638	68.5	0.016
RPCI-4	RP4	PAC	Male, blood	All	*Mbo*I	116	105,251			889	881	95.5	0.022
RPCI-5	RP5	PAC	Male, blood	All	*Mbo*I	115	142,773			1,042	1,033	116.5	0.027
RPCI-11	RP11	BAC	Male, blood	All		178	543,797	267,931	379,773/ 243,764/ 134,110	19,405	19,145	3,165.0	0.743
				1	*Eco*RI	164	108,499						
				2	*Eco*RI	168	109,496						
				3	*Eco*RI	181	109,657						
				4	*Eco*RI	183	109,382						
				5	*Mbo*I	196	106,763						
Total of top eight libraries							1,482,502	321,312	805,320/ 484,278/ 297,997	25,580	25,241	3,903.9	0.916
Total all libraries								354,510	812,594/ 488,017/ 100,775	30,445	29,298	4,260.5	1

Who was sequenced?

In Celera project 2000-2004:

21 DNA donors, for sequencing 5 of them were chosen:

one African-American,
one Asian-Chinese,
one Hispanic-Mexican,
and two Caucasians

In current paper (2007):

1 caucasian individual (J. Graig Venter), different sequence for both chromosomes – diploid human genome sequence called within the paper as HuRef

What was done?

- Additional sequencing of JGV DNA using traditional Sanger technology
- Read coverage (raw sequence read bp per finished sequence bp) was increased from 5.3 (2004 version) to 7.5 (2007 version). Raw reads were assembled to scaffolds. Scaffolds were mapped to human_36, where possible.
- Celera Assembler was made open-source and improved to make this work possible

General comparison with existing human genome assembly

Assembly	Assembly Subset	Number of Scaffolds	Number of Contigs	Gaps within Scaffolds	ACGT Bases	Span
NCBI Chromosomes	N/A	279	N/A	N/A	2,858,012,806	3,080,419,480
NCBI All	N/A	367	N/A	N/A	2,870,607,502	3,093,104,542
WGSA Chromosomes	N/A	4,940	211,493	206,553	2,659,468,408	2,993,154,503
HuRef Assembly	Chromosomes	1,408	66,762	66,354	2,782,357,138	2,809,547,336
	Scaffolds ≥ 100 kb	553	65,932	65,379	2,779,929,229	2,806,091,853
	Scaffolds ≥ 3 kb	4,528	71,343	66,815	2,809,774,459	2,844,046,670
	All scaffolds	188,394	255,300	66,906	3,002,932,476	3,037,726,076

doi:10.1371/journal.pbio.0050254.t002

98% of human_36 covered by HuRef alignments
59% of ChrY covered
95% ChrX covered by alignments, but alignments contain more gaps

150 Mb HuRef cannot be aligned
9 Mb of this (0.3% of HuRef) fills the human_36 gaps

General comparison with human_36

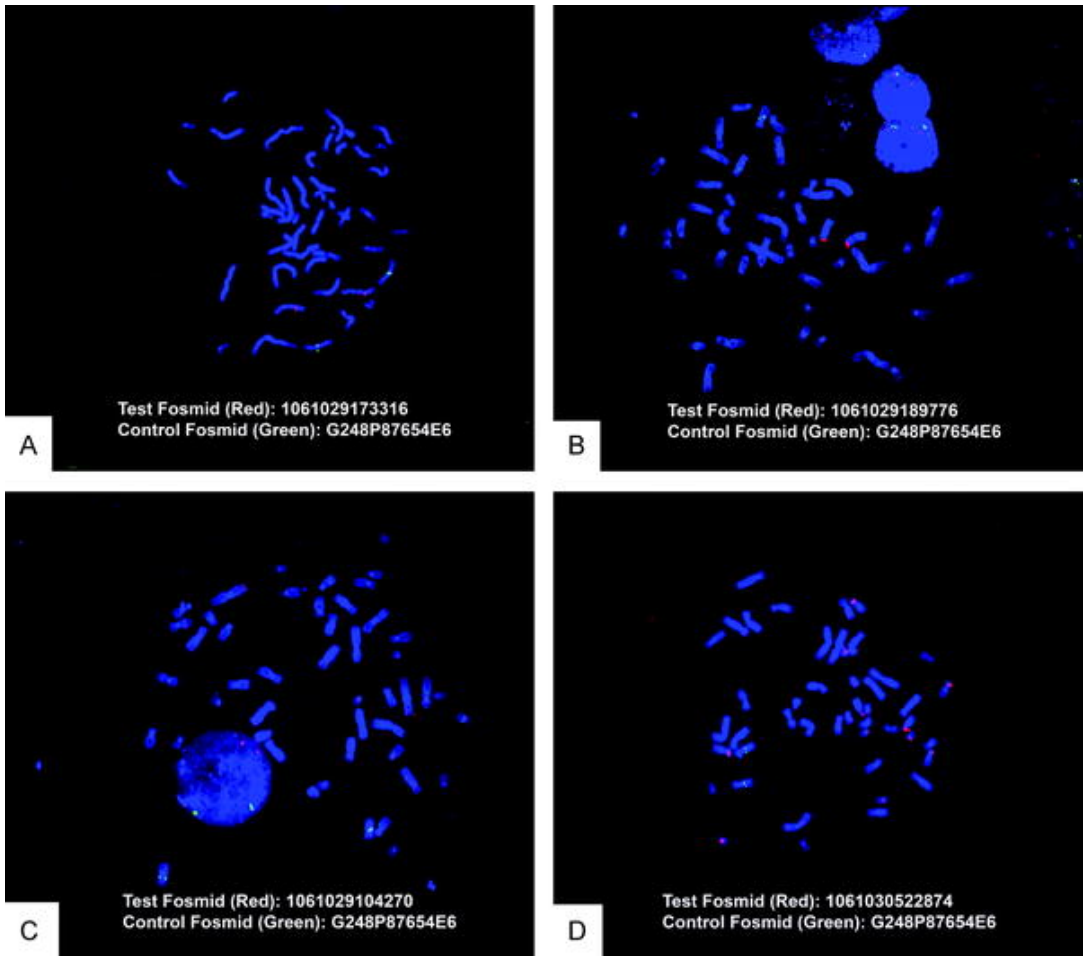


Figure 10. Non-Mapped HuRef Sequences Mapped to Coriell DNA Samples by FISH

Sequences from the HuRef donor that had no match based on the one-to-one mapping or BLAST when compared to the NCBI Human reference genome were tested by FISH. Fosmids were used as probes and the experiments were run, using Coriell DNA, to confirm the localization of the contigs or to map contigs with no prior mapping information. Shown here are four representative results. (A) An insertion at 7q22 where the FISH confirmed the HuRef mapping, (B) FISH result confirming the mapping of a sequence extending into a gap at 1p21. (C) Localization of a contig with no prior mapping information to chromosomal band 1q42. (D) An example of euchromatic-like sequence with no prior mapping information, which hybridizes to multiple centromeric locations.

DNA Variation types

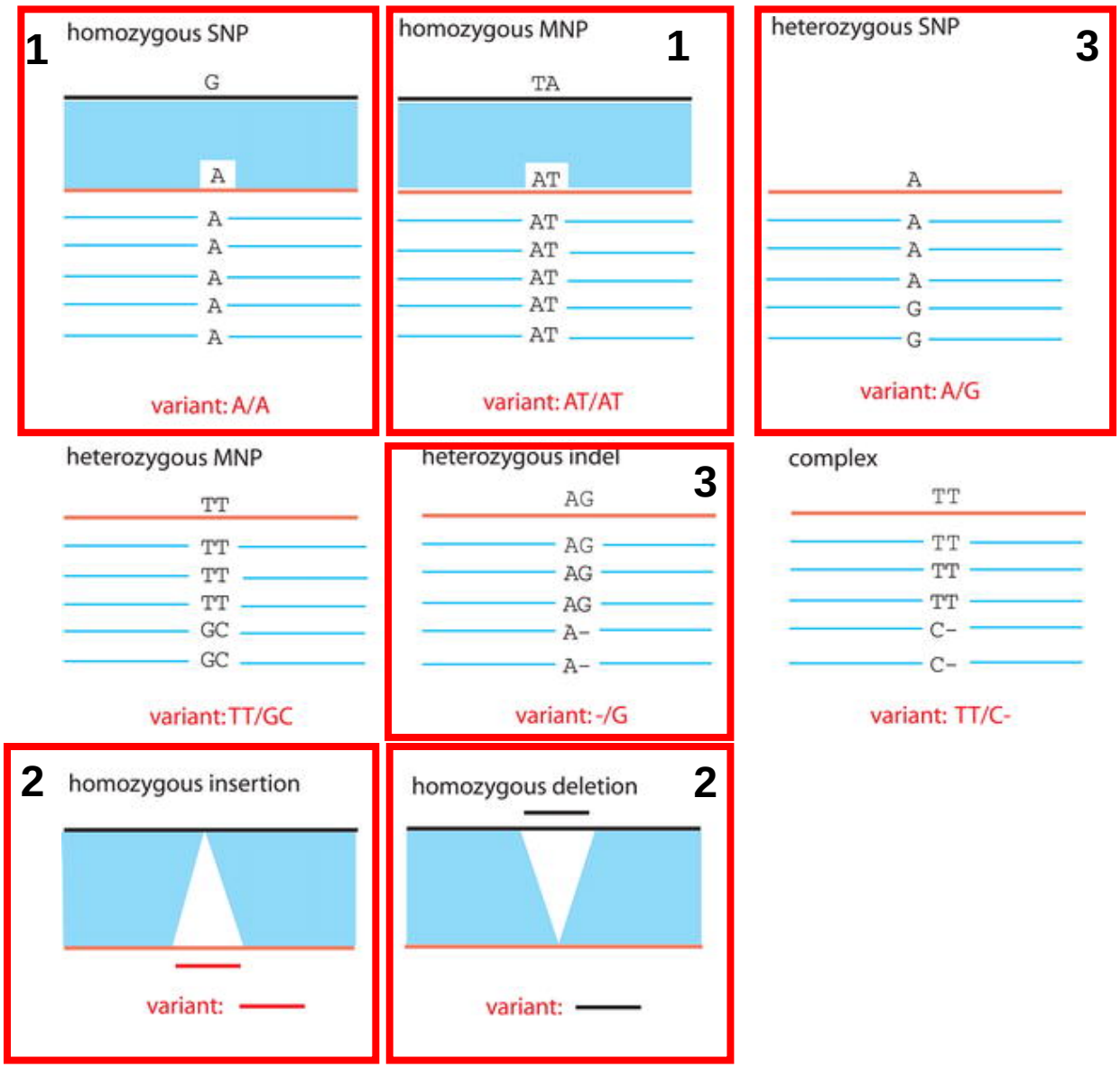


Figure 4. The Different Variant Types Identified from the HuRef Assembly and the HuRef-NCBI Assembly-to-Assembly Mapping

HuRef consensus sequence (in red) with underlying sequence reads (in blue). Homozygous variants are identified by comparing the HuRef assembly with NCBI reference assembly. Heterozygous variants are identified by base differences between sequence reads. SNP = single nucleotide polymorphism; MNP = multi-nucleotide polymorphism, which contains contiguous mismatches.

Heterozygous = variable in >20% reads

Computational Identification of DNA Variants

1. Ca 5 000 000 SNP and small indels variations discovered by comparing human_36 and HuRef but after quality control filtering only 3.3 million variants were reported
2. 690 000 large homozygous insertions, deletions, inversions (90)
3. From unaligned regions of HuRef 233 000 SNPs and small indels

Total 4.1 million variants (31% previously known).

They cover 12.3 Mb (0.4% of the genome)

Computational Identification of DNA Variants

Total 4.1 million variants (31% previously known).
They cover 12.3 Mb (0.4% of the genome)

Variant	Internal HuRef-NCBI Map	External HuRef-NCBI Map
heterozygous SNP	1,623,826	138,715
homozygous SNP	1,450,860	—
heterozygous MNP	11,825	27,160
homozygous MNP	14,838	—
heterozygous indel	218,301	45,622
complex	5,880	22,299
homozygous insertion	—	275,512
homozygous deletion	—	283,961
inversion	—	90
Total	3,325,530	798,359

By definition, homozygous insertion/deletion polymorphisms are not in regions of HuRef that align to NCBI

doi:10.1371/journal.pbio.0050254.t004

Characterization of DNA Variants

657 insertions and 659 deletions with length >100bp are related to SINEs (mostly Alu family of repeats).

90% of these belong to the youngest AluY, mostly unidentified variants of AluY

Experimental Validation of Variants (SNPs)

Presence of SNPs was validated by using modern microarray platforms:

- Affymetrix 500k (twice)
- Illumina HumanHap 650Y

Discordant results between experiments (ca 0.15%) were excluded from further analysis (array-related technological problems).1 030 000 genotype calls were compared to the variants discovered from HuRef sequence.

8.4% (86 157) of genotypes were different !!!

7% of these are heterozygous on microarray but homozygous on sequence (7.5x coverage is not sufficient to detect all variants).

1.4% of these were different homozygotes (errors?).

0.04% were homo->hetero or hetero->hetero changes

Method	Affymetrix/Illumina	Homozygous (HuRef)	Heterozygous (HuRef)	Total	Total Overlap
Affymetrix	Homozygous	4,886 (13.98%)	245 (0.70%)	34,960	468,109
	Heterozygous	29,826 (85.31%)	3 (0.01%)		
Illumina	Homozygous	7,093 (13.09%)	56 (0.10%)	54,183	649,334
	Heterozygous	46,892 (86.54%)	142 (0.26%)		
Non-redundant	Homozygous	14,035 (16.29%)	253 (0.29%)	86,157	1,029,688
	Heterozygous	71,673 (83.89%)	145 (0.17%)		

doi:10.1371/journal.pbio.0050254.t009

Experimental Validation of Variants (heterozygous indels)

Presence of indels was validated by using PCR:

19 non-genic heterozygous indels, length 1 - 16 bp

3 Coriell DNA samples and HuRef were analyzed by PCR

15 successful PCRs

4 cases: indel was identified in all 4 DNAs (common variant)

3 cases: indel was identified in HuRef only (rare variant)

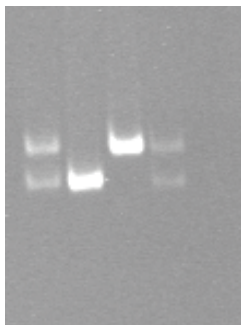


Figure S1. PAGE detection of an 8-bp indel (GATAAGTG/--------) in three Coriell DNA samples (lane 1 = NA05392 , lane 2 = NA05398, lane 3 = NA07752, and lane 4 = HuRef donor DNA).
Note the detection of two bands signifying the presence of two allelic forms in individual NA05392 and HuRef and the short and long alleles in individuals NA05398 and NA07752 respectively.

Experimental Validation of Variants (homozygous indels)

Presence of indels was validated by using PCR:

51 homozygous indels, length 100 - 1000 bp

93 Coriell DNA samples and HuRef were analyzed by PCR

43 successful PCRs

22 cases: indel was homozygous in HuRef

14 cases: indel was heterozygous in HuRef (sequencing missed other allele)

4 cases (10% !!!): there was no indel in any DNA (False Positive)

3 cases: there was no indel in HuRef but other DNAs had it

Experimental Validation of Variants (Copy Number Variants)

Presence of CNVs was validated by microarrays:

Agilent 244k (CGH Analytics)

Nimblegen 385k (CNVfinder)

Affymetrix 500k (3 different algorithms: dChip, CNAG, GEMCA)

Illumina 650Y

Most results were compared to single reference DNA. dChip, CNAG and Illumina data were compared to panel of normal individuals. All previously known CNVs were removed from reference samples.

62 CNVs (32 losses and 30 gains) were detected.

The majority of CNVs were detected by single platform only!!!

87% of the variants have been previously described in Database of Genomic Variants

5% could be false positive according to selected thresholds

No attempt was made to find CNVs from sequence alone!!!

Haplotype Assembly

Autosomal heterozygous variants (1.85 million variants) were assembled into haplotypes.

HuRef haplotypes were *strongly consistent* with HapMap haplotypes

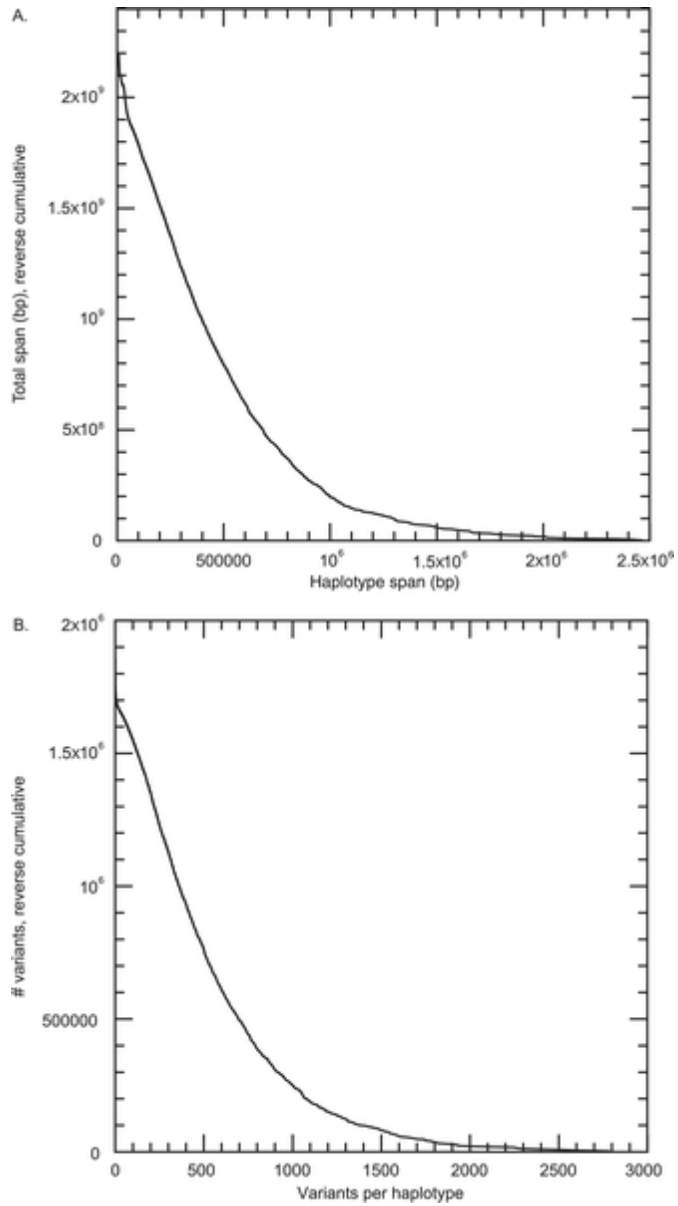


Figure 12. Distribution of Inferred Haplotype Sizes (A) Reverse cumulative distribution of haplotype spans (bp) (N50 350 kb). (B) Reverse cumulative distribution of variants per haplotype (N50 400 variants).

Tools

- Mapping reads to genomes was done with **Snapper** (http://sourceforge.net/projects/kmer/)

Conclusions

1. Human genome contains more variations than expected (0.3%-0.5% nucleotides between any two chromosomes are different).

2. Variation detection from sequence reads needs higher coverage. In this study at least 7% SNPs and 30% of long indels were missed.

3. Resequencing could not identify neither presence or location of CNVs.

4. Alu sequences are jumping around the genome (1000 such cases found).