Disclaimer

This material is only for internal use and is given to the students to prepare the final evaluation for the course of Applied Genomics of the Master degree course of Bioinformatics.

This file and its content are confidential and intended solely for the use of the individuals to whom they are given. If you have received this file it means that you are a student of the course of Applied Genomics of the master degree course in Bioinformatics, regularly enrolled for the academic year 2013-2014. If you are not a student of this course you should not disseminate, distribute or copy this file. Please notify the professor immediately by e-mail if you have received this file.
In any case, also students of this course are notified that disclosing, copying, distributing or taking any action in reliance on the contents of this information is strictly prohibited.

For the students: please note that the content of this file is not enough to pass the exam. The content of this file could contain a few errors as it has not been peer reviewed or edited after its preparation.

Applied Genomics

Program of the course:

1) Foundational concepts in genetics (including population and quantitative genetics) and genomics.
2) Genome structure and variability in vertebrates
3) The transcriptional landscape of the mammalian genome
4) High throughput technologies for genotyping and next generation sequencing (NGS) platforms
5) Applications of NGS, array comparative genome hybridization
6) Linkage disequilibrium and linkage analysis, genetic mapping
7) QTL mapping, eQTL
8) Candidate gene analysis, genome wide association studies, selection signature
9) Relevant genomic projects: modENCODE, ENCODE, 1000 genome project, The Mammalian Genome Project, 10K Genome project.
10) Discussion of relevant scientific literature

BAM files


9694451 9694461 9694471 9694481 9694491 9694501
CTCCTTACCCTCCTACCAATCCACTTCCTTCTTGCACTCCTCAGGTGCCCCCAATTCTCCAGATGCC
.....Y.....Y...Y....YR.....
ctccttacc CCTTCTTGCACTCCTCAGGTGCCCCCAATTCTCCAGATGCC
ctccttaccctcctaccaatccacttccttcttgcattcctcaggcgcccccaattctccagatgcc
ctccttaccctcctaccaatccacttccttcttgcattcctcaggcgcccccaattctccagatgcc
ctccttaccctcctaccaatctacttccttcttgcactccccaggcacccccaattctccag
ctccttaccctcctaccaatctacttccttcttgcactccccaggcacccccaattctccag
ctccttaccctcctaccaatccacttccttcttgcattcctcaggcgcccccaattctcc
ctccttaccctcctaccaatccacttccttcttgcattcctcaggcgcccccaattctcc
ctccttaccctcctaccaatccacttccttcttgcattcctcaggcg*ccccaattctcc
ctccttaccctcctaccaatccacttccttcttgcactcctcag
ttaccctcctaccaatccacttccttcttgcactcctcaggcacccccaattctccagatgcc
CCTTCTTGCACTCCTCAGGTGCCCCCAATTCTCCAGATGCC

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2012, Article ID 251364, 11 pages
doi:10.1155/2012/251364

Review Article

Comparison of Next-Generation Sequencing Systems

Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law

NGS Sequencing Department, Beijing Genomics Institute (BGI), 4th Floor, Building 11, Beishan Industrial Zone, Yantian District, Guangdong, Shenzhen 518083, China

Table 1: (a) Advantage and mechanism of sequencers. (b) Components and cost of sequencers. (c) Application of sequencers.

(a)

Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400~900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput

(b)

Sequencers	454 GS FLX	HiSeq 2000	SOLiDv4	3730xl
Instrument price	Instrument $500,000, $7000 per run	Instrument $690,000, $6000/(30x) human genome	Instrument $495,000, $15,000/100 Gb	Instrument $95,000, about $4 per 800 bp reaction
CPU	2* Intel Xeon X5675	2* Intel Xeon X5560	8* processor 2.0 GHz	Pentium IV 3.0 GHz
Memory	48 GB	48 GB	16 GB	1 GB
Hard disk	1.1 TB	3 TB	10 TB	280 GB
Automation in library preparation	Yes	Yes	Yes	No
Other required device	REM e system	cBot system	EZ beads system	No
Cost/million bases	$10	$0.07	$0.13	$2400

(c)

Sequencers	454 GS FLX	HiSeq 2000	SOLiDv4	3730xl
Resequencing		Yes	Yes	
De novo	Yes	Yes		Yes
Cancer	Yes	Yes	Yes	
Array	Yes	Yes	Yes	Yes
High GC sample	Yes	Yes	Yes	
Bacterial	Yes	Yes	Yes	
Large genome	Yes	Yes		
Mutation detection	Yes	Yes	Yes	Yes

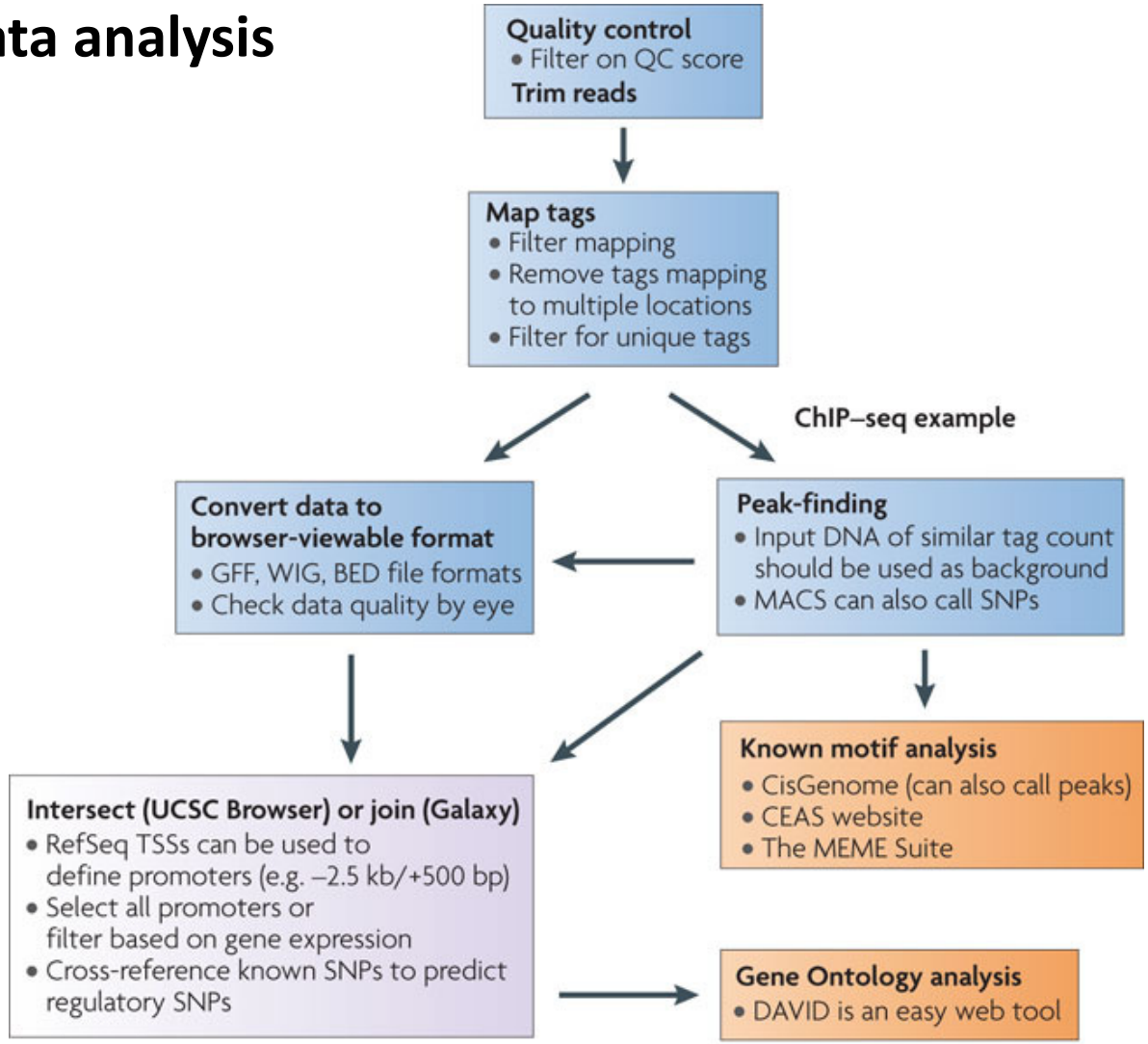
(1) All the data is taken from daily average performance runs in BGI. The average daily sequence data output is about 8 Tb in BGI when about 80% sequencers (mainly HiSeq 2000) are running.
(2) The reagent cost of 454 GS FLX Titanium is calculated based on the sequencing of 400 bp; the reagent cost of HiSeq 2000 is calculated based on the sequencing of 200 bp; the reagent cost of SOLiDv4 is calculated based on the sequencing of 85 bp.
(3) HiSeq 2000 is more flexible in sequencing types like 50SE, 50PE, or 101PE.
(4) SOLiD has high accuracy especially when coverage is more than 30x, so it is widely used in detecting variations in resequencing, targeted resequencing, and transcriptome sequencing. Lanes can be independently run to reduce cost.

Table 1 Applications of next-generation DNA sequencing

Method	Sequencing to determine:	Example reference	'Subway' route as defined in Figure 3
DNA-Seq	A genome sequence	57	Comparison, 'anatomic' (isolation by anatomic site), flow cytometery, DNA extraction, mechanical shearing, adaptor ligation, PCR and sequencing
Targeted DNA-Seq	A subset of a genome (for example, an exome)	20	Comparison, cell culture, DNA extraction, mechanical shearing, adaptor ligation, PCR, hybridization capture, PCR and sequencing
Methyl-Seq	Sites of DNA methylation, genome-wide	34	Perturbation, genetic manipulation, cell culture, DNA extraction, mechanical shearing, adaptor ligation, bisulfite conversion, PCR and sequencing
Targeted methyl-Seq	DNA methylation in a subset of the genome	129	Comparison, cell culture, DNA extraction, bisulfite conversion, molecular inversion probe capture, circularization, PCR and sequencing
DNase-Seq, Sono-Seq and FAIRE-Seq	Active regulatory chromatin (that is, nucleosome-depleted)	113	Perturbation, cell culture, nucleus extraction, DNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing
MAINE-Seq	Histone-bound DNA (nucleosome positioning)	130	Comparison, cell culture, MNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing
ChIP-Seq	Protein-DNA interactions (using chromatin immunoprecipitation)	131	Comparison, 'anatomic', cell culture, cross-linking, mechanical shearing, immunoprecipitation, DNA extraction, adaptor ligation, PCR and sequencing
RIP-Seq, CLIP-Seq, HITS-CLIP	Protein-RNA interactions	46	Variation, cross-linking, 'anatomic', RNase digestion, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, PCR and sequencing
RNA-Seq	RNA (that is, the transcriptome)	39	Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing
FRT-Seq	Amplification-free, strand-specific transcriptome sequencing	119	Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, adaptor ligation, reverse transcription and sequencing
NET-Seq	Nascent transcription	41	Perturbation, genetic manipulation, cell culture, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, circularization, PCR and sequencing
Hi-C	Three-dimensional genome structure	71	Comparison, cell culture, cross-linking, proximity ligation, mechanical shearing, affinity purification, adaptor ligation, PCR and sequencing
Chia-PET	Long-range interactions mediated by a protein	73	Perturbation, cell culture, cross-linking, mechanical shearing, immunoprecipitation, proximity ligation, affinity purification, adaptor ligation, PCR and sequencing
Ribo-Seq	Ribosome-protected mRNA fragments (that is, active translation)	48	Comparison, cell culture, RNase digestion, ribosome purification, RNA extraction, adaptor ligation, reverse transcription, rRNA depletion, circularization, PCR and sequencing
TRAP	Genetically targeted purification of polysomal mRNAs	132	Comparison, genetic manipulation, 'anatomic', cross-linking, affinity purification, RNA extraction, poly(A) selection, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing
PARS	Parallel analysis of RNA structure	42	Comparison, cell culture, RNA extraction, poly(A) selection, RNase digestion, chemical fragmentation, adaptor ligation, reverse transcription, PCR and sequencing
Synthetic saturation mutagenesis	Functional consequences of genetic variation	93	Variation, genetic manipulation, barcoding, RNA extraction, reverse transcription, PCR and sequencing
Immuno-Seq	The B-cell and T-cell repertoires	86	Perturbation, 'anatomic', DNA extraction, PCR and sequencing
Deep protein mutagenesis	Protein binding activity of synthetic peptide libraries or variants	95	Variation, genetic manipulation, phage display, *in vitro* competitive binding, DNA extraction, PCR and sequencing
PhIT-Seq	Relative fitness of cells containing disruptive insertions in diverse genes	92	Variation, genetic manipulation, cell culture, competitive growth, linear amplification, adaptor ligation, PCR and sequencing

FAIRE-seq, formaldehyde-assisted isolation of regulatory elements–sequencing. MAINE-Seq, MNase-assisted isolation of nucleosomes-sequencing; RIP-Seq, RNA-binding protein immunoprecipitation-sequencing; CLIP-Seq, cross-linking immunoprecipitation-sequencing; HITS-CLIP, high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation; FRT-Seq, on-flowcell reverse transcription–sequencing. NET-Seq, native elongating transcript sequencing. TRAP, translating ribosome affinity purification. PhIT-Seq, phenotypic interrogation via tag sequencing.

Flow chart for data analysis



tion	SIFT	http://blocks.fhcrc.org/sift/SIFT.html	
	Po	d.edu/pph2	
	XVAR	http://xvar.org	
	CHASM		
	CIRCOS	http://mkweb.bcgsc.ca/circos	Essentially all papers use CIRCOS to display genomic events
	IGV	http://www.broadinstitute.org/igv	IGV is used to display genomic events and for manual review

Computational tools 1

gnment methods with a brief description of each is constantly updated at http://en.wikipedia.org/wiki/List_of_sequence_alignment_softw

Computational tools 2

Scope	Program	Repeat-relevant parameters	Website	Refs
SV or CNV detection	BreakDancer	Specify the mapping quality threshold for ambiguous reads: -q	http://sourceforge.net/projects/breakdancer	
	CNVnator	None available or none required	http://sv.gersteinlab.org/cnvnator/	
	He *et al.* (2011)	Algorithm only, able to estimate CNV counts in repeat-rich regions	None	47
	PEMer	Maximum alignments per multi-read: --max_duplicates_per_score	http://sv.gersteinlab.org/pemer	
	VariationHunter	None available or none required	http://compbio.cs.sfu.ca/strvar.htm	
SNP detection	GATK	None available or none required	http://www.broadinstitute.org/gsa/wiki/index.php/Downloading_the_GATK	
	SAMtools	In repetitive regions, avoid calling 'A': -avcf ref.fa aln.bam	http://samtools.sourceforge.net	
	SOAPsnp	None required; multi-reads supported by read aligner parameters	http://soap.genomics.org.cn/soapsnp.html	
	Sniper	Read mapping policy: --all, --uniq, --best	http://kim.bio.upenn.edu/software/sniper.shtml	
	VarScan	None available or none required	http://varscan.sourceforge.net	
Short-read alignment	Bowtie	Randomly distribute reads across repeats: --best –M 1 -strata	http://bowtie-bio.sourceforge.net	16
	BFAST	Reports all locations by default	http://bfast.sourceforge.net	69
	Burrows–Wheeler Aligner (BWA)	Report one random hit for repetitive reads: -n 1	http://bio-bwa.sourceforge.net	70
	mrFAST	Reports all locations by default, for best match: --best	http://mrfast.sourceforge.net	71
	SOAPAligner	Report all locations: -r 2	http://soap.genomics.org.cn/soapaligner.html	72
De novo assembly	Allpaths-LG	None required: incorporated into library insert size recipe	http://www.broadinstitute.org/software/allpaths-lg/blog/?page_id=12	20
	CABOG	Re-assemble misclassified non-unique unitigs: doToggle=1	http://wgs-assembler.sf.net	73
	SGA	Resolve small repeats at end of reads: -r 20	http://github.com/jts/sga	
	SOAPdenovo	Use reads to solve small repeats: -R	http://soap.genomics.org.cn/soapdenovo.html	17
	Velvet	Use long reads to resolve repeats: -long, -exp_cov auto	http://www.ebi.ac.uk/~zerbino/velvet	74,75

NGS

OPEN ACCESS Freely available online

PLoS one

Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology

August 2009 | Volume 4 | Issue 8 | e6524

Ramos et al.

Animals and DNA

DNA samples were obtained from five pig breeds, including Duroc (DU), Pietrain (PI), Landrace (LR), Large White (LW) and Wild Boar (WB).

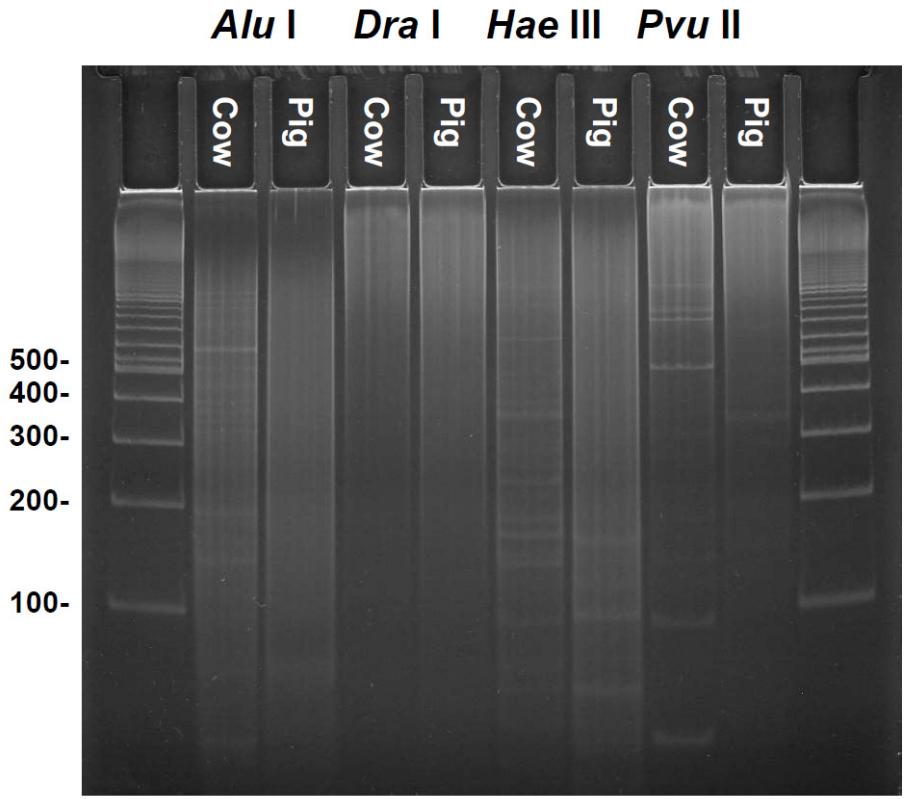
For each breed, a single DNA pool containing equal DNA amounts from all of the samples was prepared.

The number of animals per pool was 34, 23, 29, 36 and 36 for DU, PI, LR, LW and WB, respectively.

The DNA samples for the commercial breeds were representative of the worldwide distribution of the breeds (samples originated from the USA, Netherlands and Denmark), whereas the WB samples were collected mainly in Europe, with five samples originating from Japan.

Construction of the reduced representation libraries

For each breed, a total of 25 µg from each DNA pool, divided in five aliquots of 5 µg each, was digested with each of three restriction enzymes *Alu*I, *Hae*III and *Msp*I



Problem:

Restriction Enzymes can also be classified by the numbers of bases in the recognition sequence

The numbers of bases will determine the frequency of that specific sequence in an average DNA sequence.

For example:

*Alu*I recognizes a 4 bp sequence with would occur once every

*Pvu*I recognizes a 6 bp sequence with would occur once every

*Not*I recognizes an 8 bp sequence with would occur once ever

Problem:

Restriction Enzymes can also be classified by the numbers of bases in the recognition sequence

The numbers of bases will determine the frequency of that specific sequence in an average DNA sequence.

For example:

*Alu*I recognizes a 4 bp sequence with would occur once every 4^4 or 256 bp.

*Pvu*I recognizes a 6 bp sequence with would occur once every 4^6 or 4,096 bp.

*Not*I recognizes an 8 bp sequence with would occur once every 4^8 or 65,536 bp.

The restriction enzyme HindIII recognizes the sequence AAGCTT.
If genomic DNA of random sequence is cleaved with HindIII, what will be the average size of the the fragment produced ?

The restriction enzyme HindIII recognizes the sequence AAGCTT.
If genomic DNA of random sequence is cleaved with HindIII, what will be the average size of the the fragment produced ?

The chance that any one base, A, C, G, or T, will occur at any particular position in DNA of random sequence is one in four. The number of cutting sites in random DNA can be estimated by raising ¼ to n, where n is the number of bps in the recognition sequence: $((1/4)^n)$...

Then?

10kb x $((1/4)^n)$

Sequencing

All 19 libraries were sequenced on a 1G Genome Analyzer (Illumina, San Diego, CA, USA). The length of all sequences generated in this study was 36 nucleotides.

In addition, two pooled samples derived from each of the *Alu*I RRLs were prepared and sequenced using a 454 FLX system (Roche Applied Science, Indianapolis, IN, USA) on the GS FLX platform.

Processing of the GA sequences

The raw sequence data were filtered according to different criteria.

Each sequence was first evaluated for the presence of the expected sequence motif for each restriction enzyme; sequences not containing the expected sequence were discarded. The acceptable sequences began with CT, CC or CGG for the *Alu*I, *Msp*I, and *Hae*III restriction digests, respectively.

Sequences that contained the same nucleotide at more than 18 continuous positions were also eliminated.

The average quality score was next calculated for each read by averaging the individual score for each of the 36 base positions. Previous results obtained in a smaller pilot study indicated that a minimum average quality score of 12 was an acceptable threshold, and, therefore, all sequence reads with an average quality score<12 were removed from the dataset.

SNP discovery

Sequences were aligned against the assembled reference genome and initial SNP detection was performed using MAQ.

For SNP discovery, only reads that aligned to a single unique location of the genome were considered. Because MAQ calls a SNP as being any difference between the reads or between the reads and the reference genome, the initial MAQ SNP prediction output is large and must be filtered. Several criteria were used to exclude the less reliable SNPs from the dataset. Thresholds were established for a number of MAQ values that were useful in predicting reliable SNPs.

Specifically, MAQ's minimal map quality for the read, minimal consensus quality and minimal map quality of the best mapping read for each predicted SNP position were used as criteria to select reliable SNPs by setting the thresholds for all three parameters at 10 (SNPs with any values<10 were discarded).

Moreover, the minor allele at each SNP needed to be represented in at least three reads and that the total number of reads per SNP was lower than 120 (SNPs with higher read depth were discarded).

Maq: Mapping and Assembly with Qualities

Introductions

Maq is a software that builds mapping assemblies from short reads generated by the next-generation sequencing machines. It is particularly designed for Illumina-Solexa 1G Genetic Analyzer, and has preliminary functions to handle ABI SOLiD data.

Maq first aligns reads to reference sequences and then calls the consensus. At the mapping stage, maq performs **ungapped** alignment. For single-end reads, maq is able to find all hits with up to 2 or 3 mismatches, depending on a command-line option; for paired-end reads, it always finds all paired hits with one of the two reads containing up to 1 mismatch. At the assembling stage, maq calls the consensus based on a statistical model. It calls the base which maximizes the posterior probability and calculates a phred quality at each position along the consensus. Heterozygotes are also called in this process.

Get Maq

Maq is distributed under GNU Public License (GPL). All its source codes are freely available to both academic and commercial users. The latest version can be downloaded at the SourceForge download page. Source codes are also available at the SourceForge subversion server, which can be accessed with:

```
svn co https://mapass.sourceforge.net/svnroot/mapass/trunk/maq maq
```

Install Maq

There are two ways to compile maq. The first way is to use the GNU building systems. Simply type:

```
./configure; make; make install
```

you can compile and install maq. Three executables, `maq', `maq.pl' and `farm-run.pl', will be copied to `/usr/local/bin' by default.

Alternatively, you could compile with:

```
make -f Makefile.generic
```

Summary of the SNP discovered from the four analyzed RRLs

	AluI short	*AluI* long	*HaeIII*	*MspI*	Total
Initial MAQ output	2,625,323	2,854,329	2,377,571	1,180,640	9,037,863
Filtered SNP output	106,456	124,578	56,817	27,279	315,130
Low MAF SNPs[1]	11,149	39,096	5,620	1,891	57,756
Total High confidence SNPs	117,605	163,674	62,437	29,170	372,886

[1]SNP detected with only two minor alleles among the sequence reads.

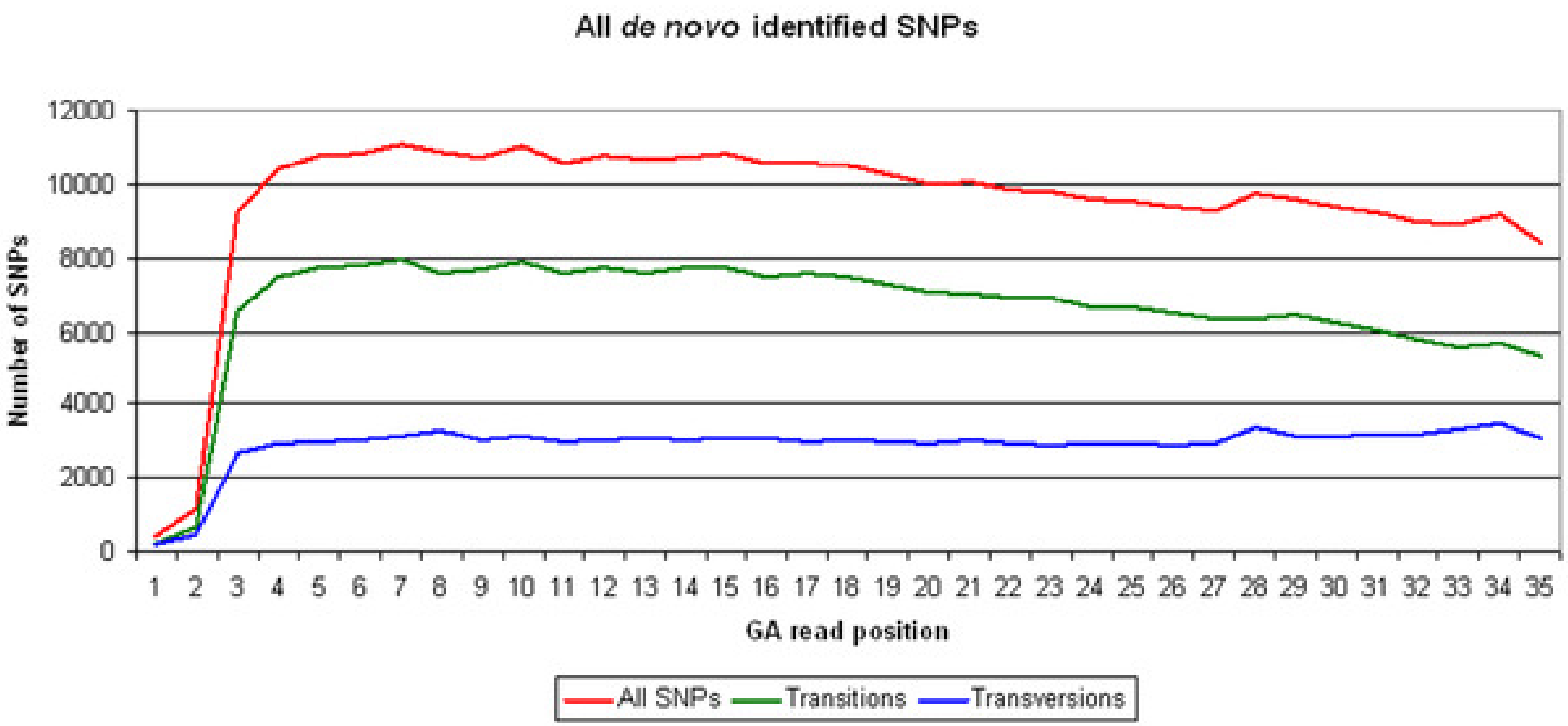
doi:10.1371/journal.pone.0006524.t002

Number of Illumina Genome Analyzer reads generated, filters applied to the dataset and final number of reads used for SNP discovery from the four RRLs

	AluI short	*AluI* long	*Hae*III	*Msp*I	Total
Starting number of reads	87,962,916	145,926,417	67,057,081	69,507,210	370,453,624
Filters	**Number of reads removed from dataset**				
Restriction enzyme motif	3,276,584	48,401,718	6,854,265	16,555,378	75,087,945
Poly-(A,C,G,T)	260,648	656,411	278,435	202,585	1,398,079
Quality score	1,004,585	1,085,621	2,138,675	402,036	4,630,917
Over-represented reads	10,167,678	14,193,925	9,559,415	7,622,490	41,543,508
Number of reads used for mapping	73,253,421	81,588,742	48,226,291	44,724,721	247,793,175
% Usable reads	83.3	55.9	71.9	64.4	66.9

doi:10.1371/journal.pone.0006524.t001

SNP distribution on each of the GA read positions



Distances between the SNPs included on the 60K+porcine Beadchip

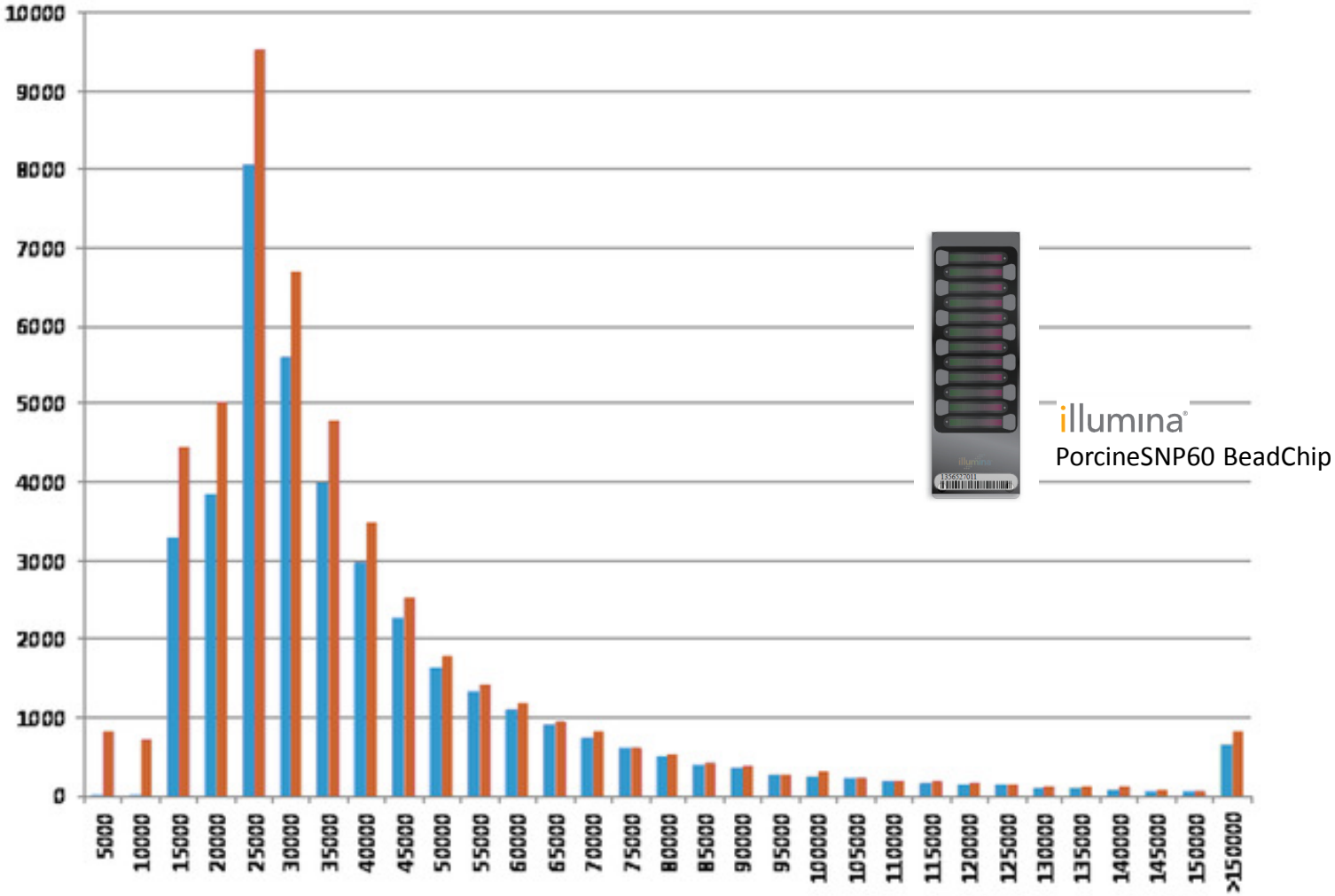
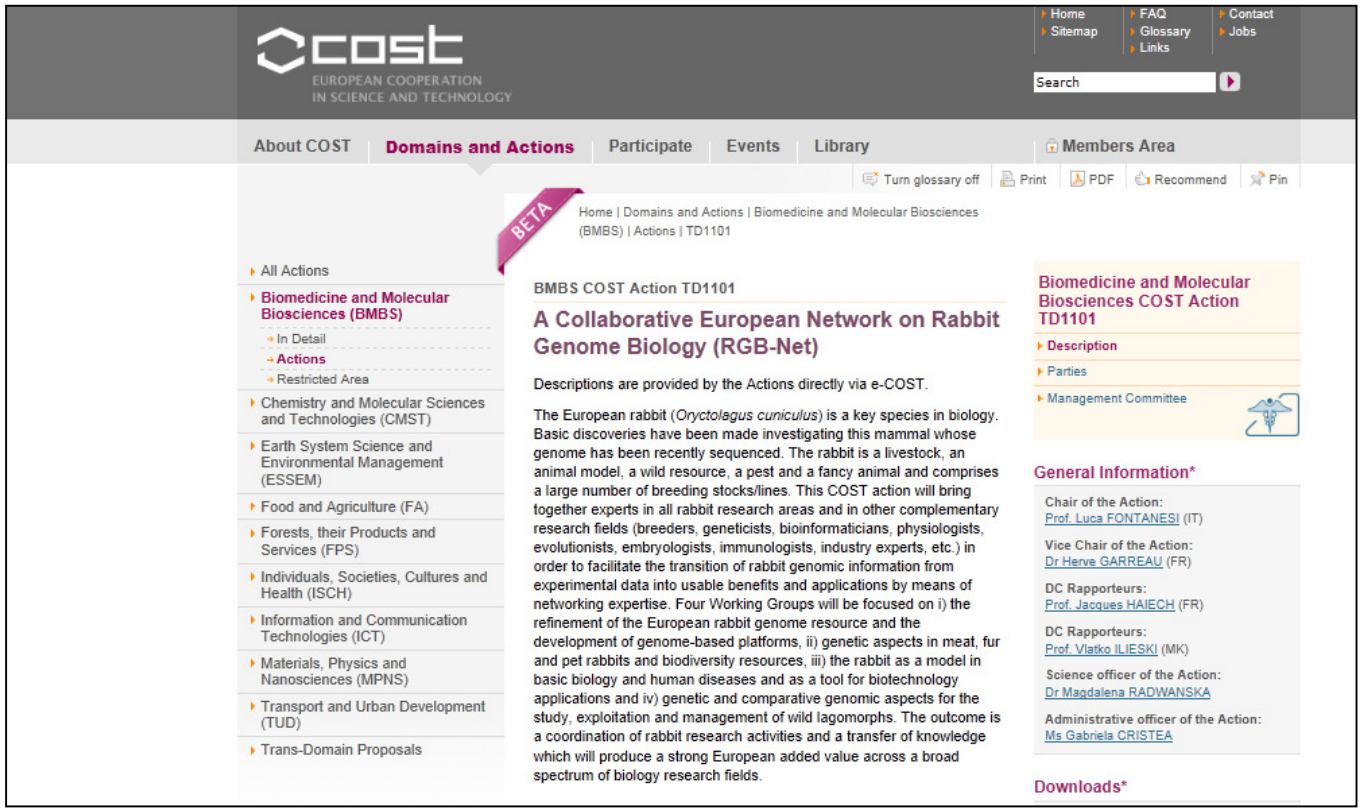


Table 2. Currently available whole-genome single nucleotide polymorphism (SNP) chips developed for important agricultural species

Species	Identification[1]	Classification	Provider[2]	Consortium	SNP, no.
Potato	Potato	Public	Illumina	SolCAP	8,303
Tomato	Tomato	Public	Illumina	SolCAP	7,720
Apple	Apple	Public	Illumina	RosBREED	8,788
Peach	Peach	Public	Illumina	RosBREED	8,144
Cherry	Cherry	Public	Illumina	RosBREED	5,696
Maize	MaizeSNP50	Commercial	Illumina	Commercial	56,110
Rice	Rice 44K	Commercial	Affymetrix	Commercial	44,100
Chicken	Chicken	Private: public sale	Illumina	Cobb Vantress-Hendrix-USDA	57,636
Cat	Feline	Private: public sale	Illumina	Morris Animal Foundation	62,897
Horse	Equine	Private: public sale	Illumina	Neogen (GeneSeek)	65,157
Sheep	Ovine	Private: public sale	Illumina	AgResearch	5,409
Cattle	BovineHD	Commercial	Illumina	Various	777,962
Cattle	BovineSNP50v2	Commercial	Illumina	Various	54,609
Cattle	BOS 1	Commercial	Affymetrix	Various	648,000
Sheep	OvineSNP50	Commercial	Illumina	Various	52,241
Cattle	BovineLD	Commercial	Illumina	Various	6,909
Pig	PorcineSNP60	Commercial	Illumina	Various	62,163
Dog	CanineHD	Commercial	Illumina	Various	173,662

[1]HD = high density; LD = low density.
[2]Illumina Inc., San Diego, CA; Affymetrix, Santa Clara, CA.

SNP discovery in the rabbit genome using reduced representation libraries



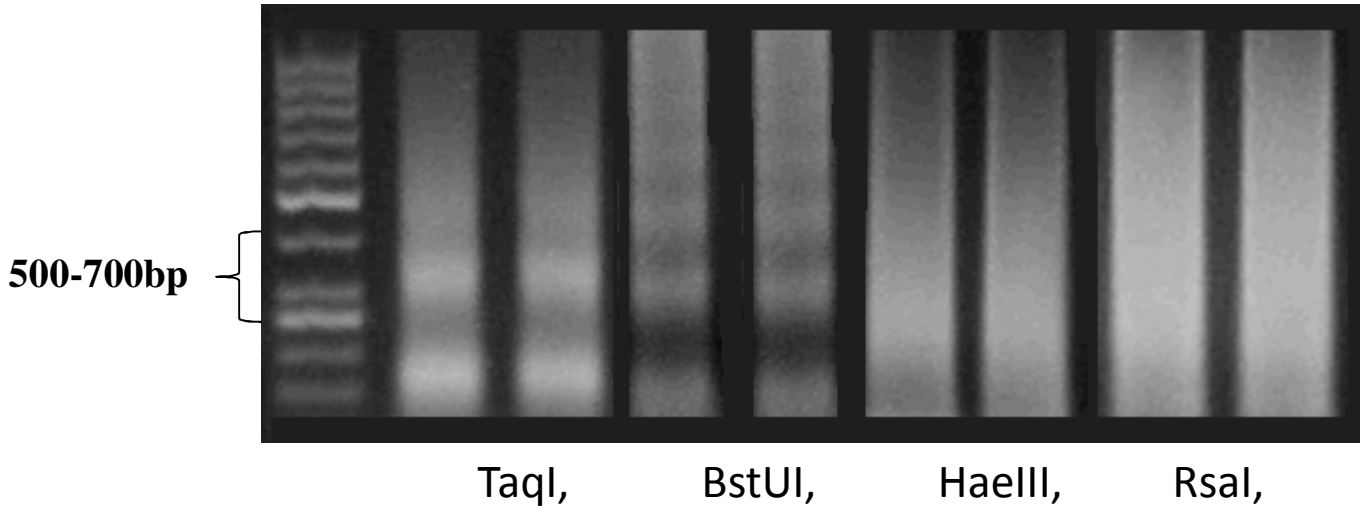
SNP discovery in the rabbit genome using reduced representation libraries

Pooled equimolar genomic DNA from 10 rabbits of different breeds

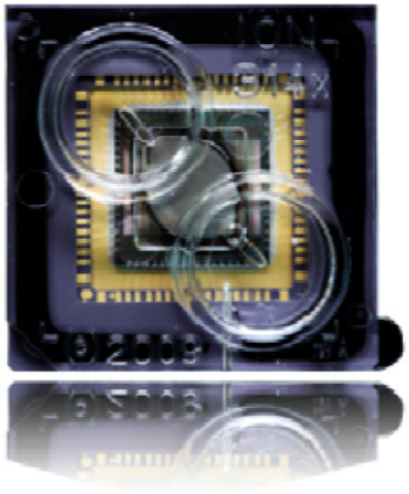
Digestion of different restriction enzymes

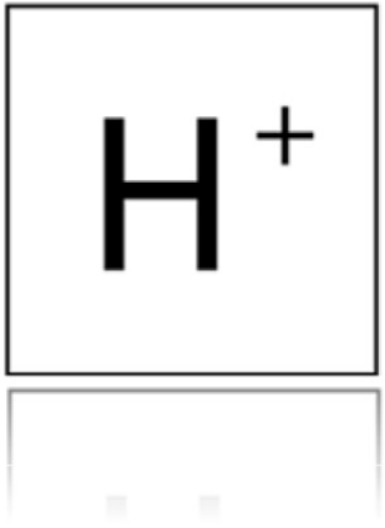
Selection of a band of 500-600 bp

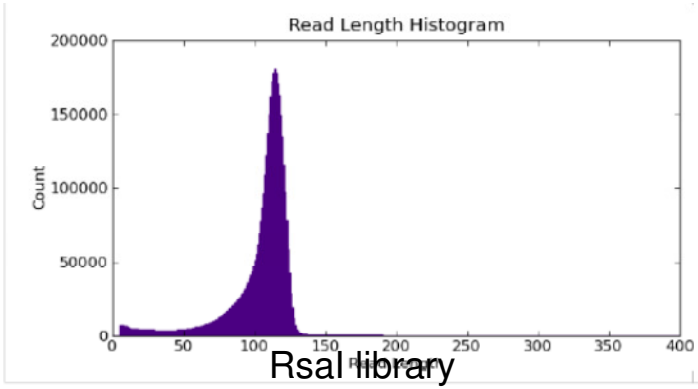
Sequencing with Ion Torrent

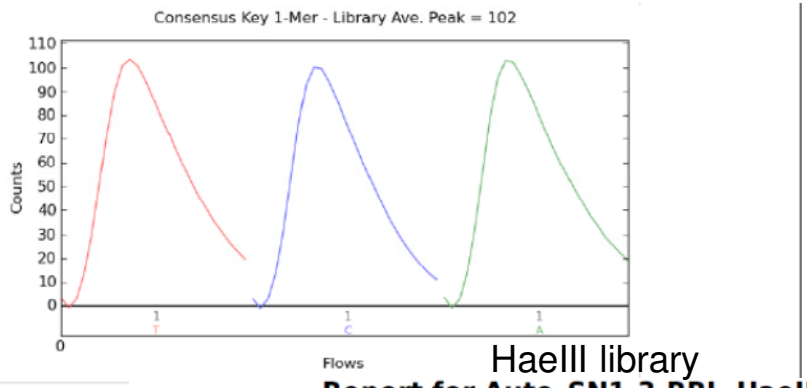


ion torrent









RsaI library

HaeIII library

Reference Genome Information

Genome Name	*Oryctolagus cuniculus*
Genome Size	2,247,769,349 bases
Genome Version	2.67
Index Version	tmap-f3

Based on Full Library Alignment to Provided Reference

	AQ20
Total Number of Bases [Mbp]	242.22
Mean Length [bp]	89
Longest Alignment [bp]	159
Mean Coverage Depth	0.10×
Percentage of Library Covered	7%

Read Alignment Distribution

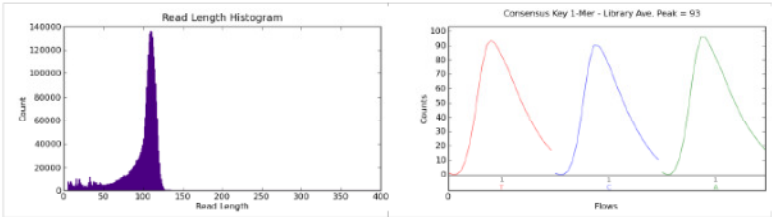
Read Length [bp]	Reads	Unmapped	Excluded	Clipped
50	3,454,467	252,356	250	0
100	2,719,053	194,657	192	199,609
150	19,164	2,283	0	16,867

Report for Auto_SN1-3-RRL_HaeIII_4

Library Summary

Based on Predicted Per-Base Quality Scores - Independent of Alignment

Total Number of Bases [Mbp]	280.51
‣ Number of Q20 Bases [Mbp]	248.49
Total Number of Reads	2,917,869
Mean Length [bp]	96
Longest Read [bp]	203



Reference Genome Information

Genome Name	*Oryctolagus cuniculus*
Genome Size	2,247,769,349 bases
Genome Version	2.67
Index Version	tmap-f3

Based on Full Library Alignment to Provided Reference

	AQ20	Perfect
Total Number of Bases [Mbp]	170.99	153.59
Mean Length [bp]	86	79
Longest Alignment [bp]	145	145
Mean Coverage Depth	0.10×	0.10×
Percentage of Library Covered	5%	5%

Read Alignment Distribution

Read Length [bp]	Reads	Unmapped	Excluded	Clipped	Perfect	1 mismatch	≥2 mismatches
50	2,397,922	167,877	196	0	1,481,975	370,928	376,946
100	1,738,566	119,282	120	113,936	743,605	317,342	444,281

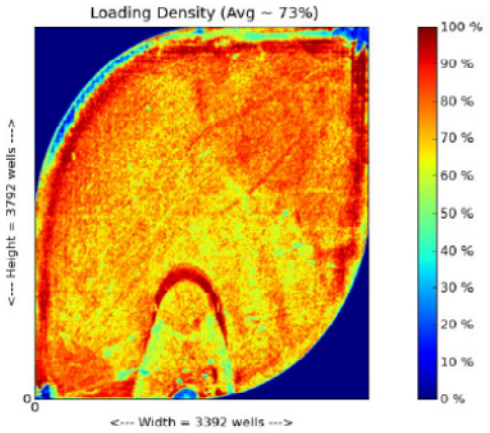
SNP discovery in the rabbit genome using reduced representation libraries

RsaI library

Ion Sphere™ Particle (ISP) Identification Summary

	Count	Percentage
Total Addressable Wells	11,302,458	
▸ Wells with ISPs	8,298,434	73%
▸ Live ISPs	8,000,872	96%
▸ Test Fragment ISPs	25,875	<1%
▸ Library ISPs	7,974,997	100%

	Count	Percentage
Library ISPs / Percent Enrichment	7,974,997	96%
▸ Filtered: Polyclonal	2,475,695	31%
▸ Filtered: Primer dimer	1,660	<1%
▸ Filtered: Low quality	1,450,771	18%
▸ **Final Library Reads**	**4,046,871**	**51%**



SNP discovery in the rabbit genome using reduced representation libraries

BAM file

SNP discovery in the rabbit genome using reduced representation libraries

Summary of the results

Rsal library: total SNP/INDEL (no filtered): 435550
total SNP/INDEL_Q10: 237555
total SNP/INDEL_Q10_(2ref-2alt): 10726

HaeIII library: total SNP/INDEL (no filtered): 305267
total SNP/INDEL_Q10 (no filetered): 154451
total SNP/INDEL_Q10_(2ref-2alt): 5480

Pooled data: total SNP/INDEL (no filtered): 730631
total SNP/INDEL_Q10: 400956
total SNP/INDEL_Q10_(2ref-2alt): 23170

frameshifts: 113; intergenic variants: 15673, 3'UTR variants: 98, cod-seq-variants: 11, downstream-gene-vaiant: 2399, intronic variants: 2780, missense variants: 203, nc_reg_variants: 31, splice-reg-var: 21, stop-gained: 5 syn-var: 229; upstream gene variant: 2034; other: 97

Vol 464|25 March 2010|doi:10.1038/nature08832 nature

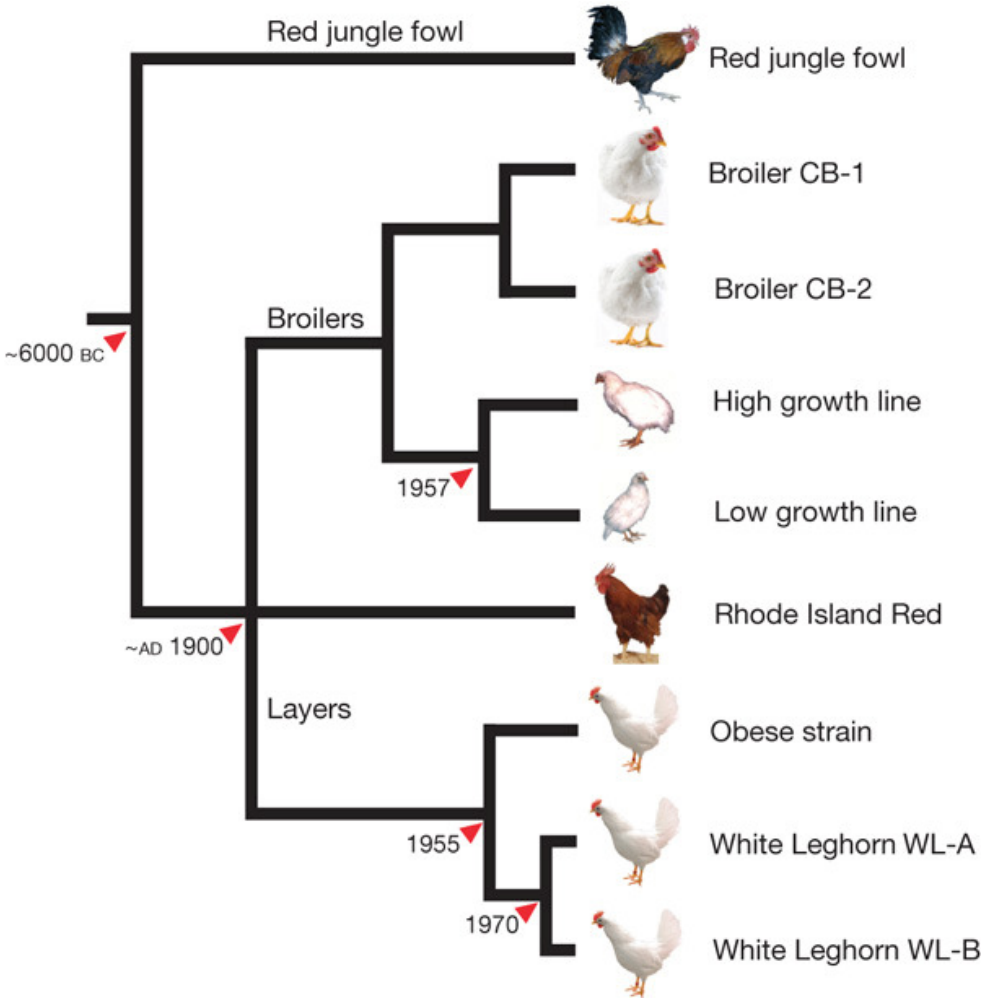
LETTERS

Whole-genome resequencing reveals loci under selection during chicken domestication

Carl-Johan Rubin[1]*, Michael C. Zody[1,2]*, Jonas Eriksson[1], Jennifer R. S. Meadows[1], Ellen Sherwood[3], Matthew T. Webster[1], Lin Jiang[1], Max Ingman[4], Ted Sharpe[2], Sojeong Ka[5], Finn Hallböök[5], Francois Besnier[6], Örjan Carlborg[6], Bertrand Bed'hom[7], Michèle Tixier-Boichard[7], Per Jensen[8], Paul Siegel[9], Kerstin Lindblad-Toh[1,2] & Leif Andersson[1,6]

A **selective sweep** is the reduction or elimination of variation among the nucleotides in neighboring DNA of a mutation as the result of recent and strong positive natural or artificial selection.

Chicken lines resequenced



AB SOLiD = 35 bp reads

The uniquely placed reads covered 92% of the 1,043 megabases (Mb)

Three independent reads of the same non-reference nucleotide were required to declare a position polymorphic.

Selective-sweep analysis

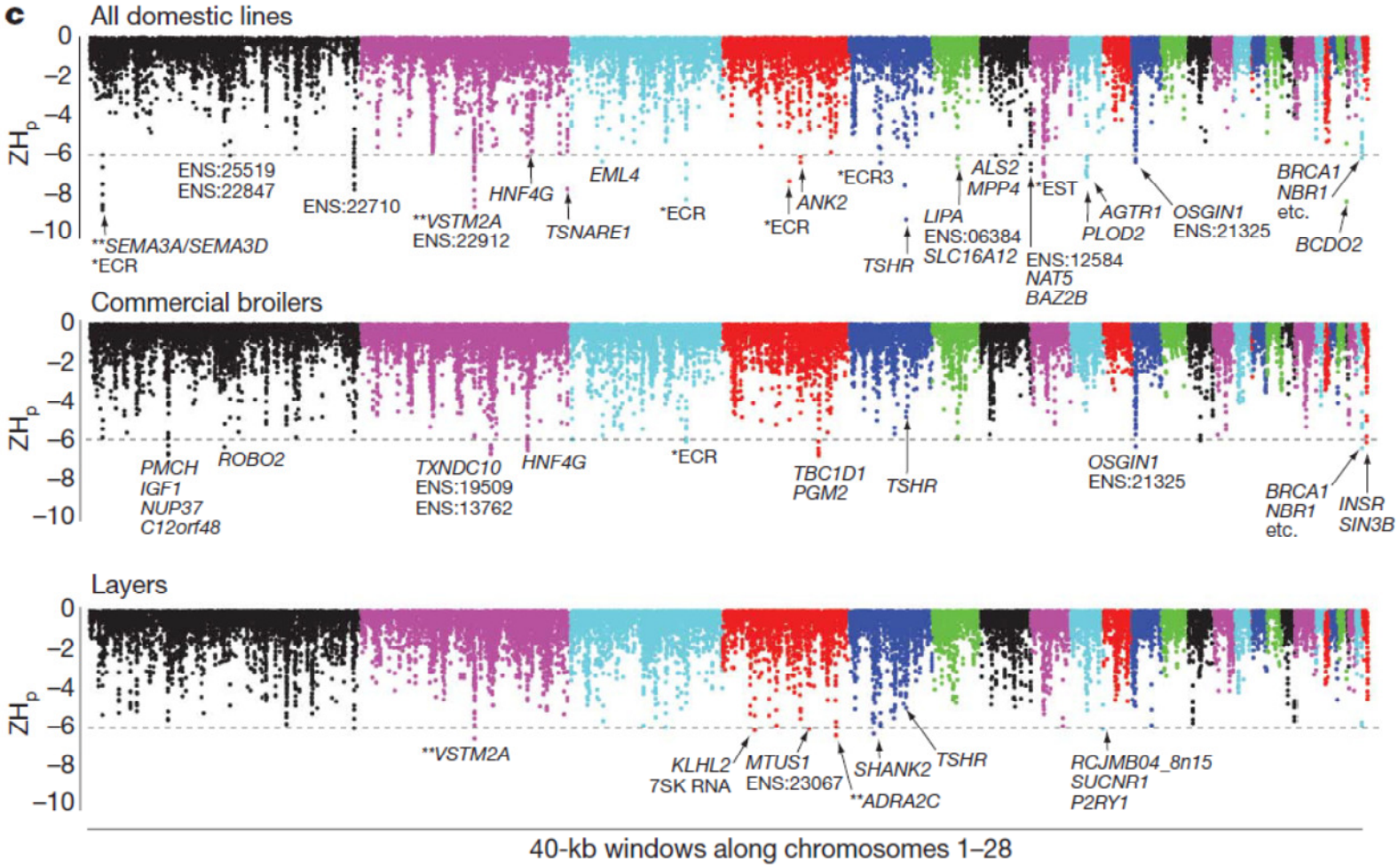
Allele counts at SNP positions are used to identify signature of selection in sliding 40-kb windows, for pools of sequence data.
For each pool and SNP, the numbers of reads corresponding to the most and least abundant allele (n_{MAJ} and n_{MIN}) are determined.
For each window in each breed pool, a pooled heterozygosity score is calculated:

$$H_p = 2\Sigma n_{MAJ}\, \Sigma n_{MIN}/(\Sigma n_{MAJ} + \Sigma n_{MIN})^2$$

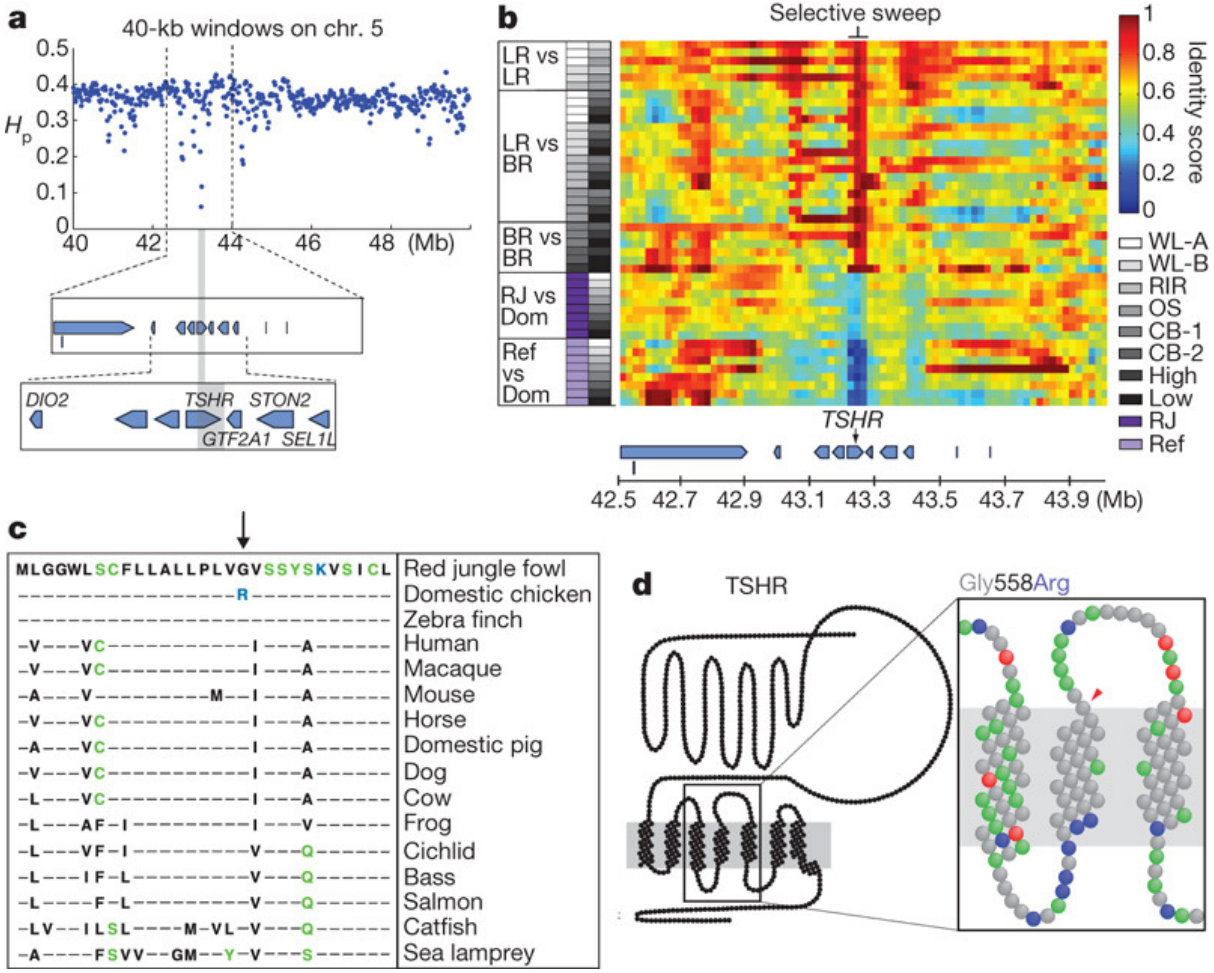
where $2\Sigma n_{MAJ}$ and Σn_{MIN} are the sums of n_{MAJ} and, respectively, n_{MIN} for all SNPs in the window.

Individual Hp values are then Z-transformed as follows:

$$ZH_p = (H_p - \mu H_p)/\sigma H_p$$


c
All domestic lines
Commercial broilers
Layers
ZHp
0
−2
−4
−6
−8
−10
**SEMA3A/SEMA3D
*ECR
ENS:25519
ENS:22847
ENS:22710
**VSTM2A
ENS:22912
HNF4G
TSNARE1
EML4
*ECR
*ECR
ANK2
*ECR3
TSHR
ALS2
MPP4
LIPA
ENS:06384
SLC16A12
*EST
ENS:12584
NAT5
BAZ2B
AGTR1
PLOD2
OSGIN1
ENS:21325
BRCA1
NBR1
etc.
BCDO2
PMCH
IGF1
NUP37
C12orf48
ROBO2
TXNDC10
ENS:19509
ENS:13762
HNF4G
*ECR
TBC1D1
PGM2
TSHR
OSGIN1
ENS:21325
BRCA1
NBR1
etc.
INSR
SIN3B
**VSTM2A
KLHL2
7SK RNA
MTUS1
ENS:23067
SHANK2
**ADRA2C
TSHR
RCJMB04_8n15
SUCNR1
P2RY1
40-kb windows along chromosomes 1–28

Details of *TSHR* mutation and sweep region.



Strong signatures of selection in the domestic pig genome

Carl-Johan Rubin[a,1], Hendrik-Jan Megens[b,1], Alvaro Martinez Barrio[a], Khurram Maqbool[c], Shumaila Sayyab[c], Doreen Schwochow[c], Chao Wang[a], Örjan Carlborg[d], Patric Jern[a], Claus B. Jørgensen[e], Alan L. Archibald[f], Merete Fredholm[e], Martien A. M. Groenen[b], and Leif Andersson[a,c,2]

[a]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, SE-751 23 Uppsala, Sweden; [b]Animal Breeding and Genomics Centre, Wageningen University, 6708 WD, Wageningen, The Netherlands; Departments of [c]Animal Breeding and Genetics and [d]Clinical Sciences, Swedish University of Agricultural Sciences, SE-750 07 Uppsala, Sweden; [e]Department of Veterinary Clinical and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, DK-1165 Copenhagen, Denmark; and [f]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian EH25 9RG, United Kingdom

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2012.

Table 1 | Assembly and annotation statistics

Assembly	Placed	Unplaced	Annotation*
Total length	2,596,639,456	211,869,922	21,640 protein-coding genes
Ungapped length	2,323,671,356	195,490,322	380 pseudogenes
Scaffolds	5,343	4,562	2,965 ncRNAs†
Contigs	73,524	168,358	197,675 gene exons
Scaffold N50	637,332	98,022	26,487 gene transcripts
Contig N50	80,720	2,423	

*Numbers refer to the annotation performed by Ensembl (release 67). Results of an independent annotation by the NCBI can be obtained from http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=9823&build=4&ver=1.

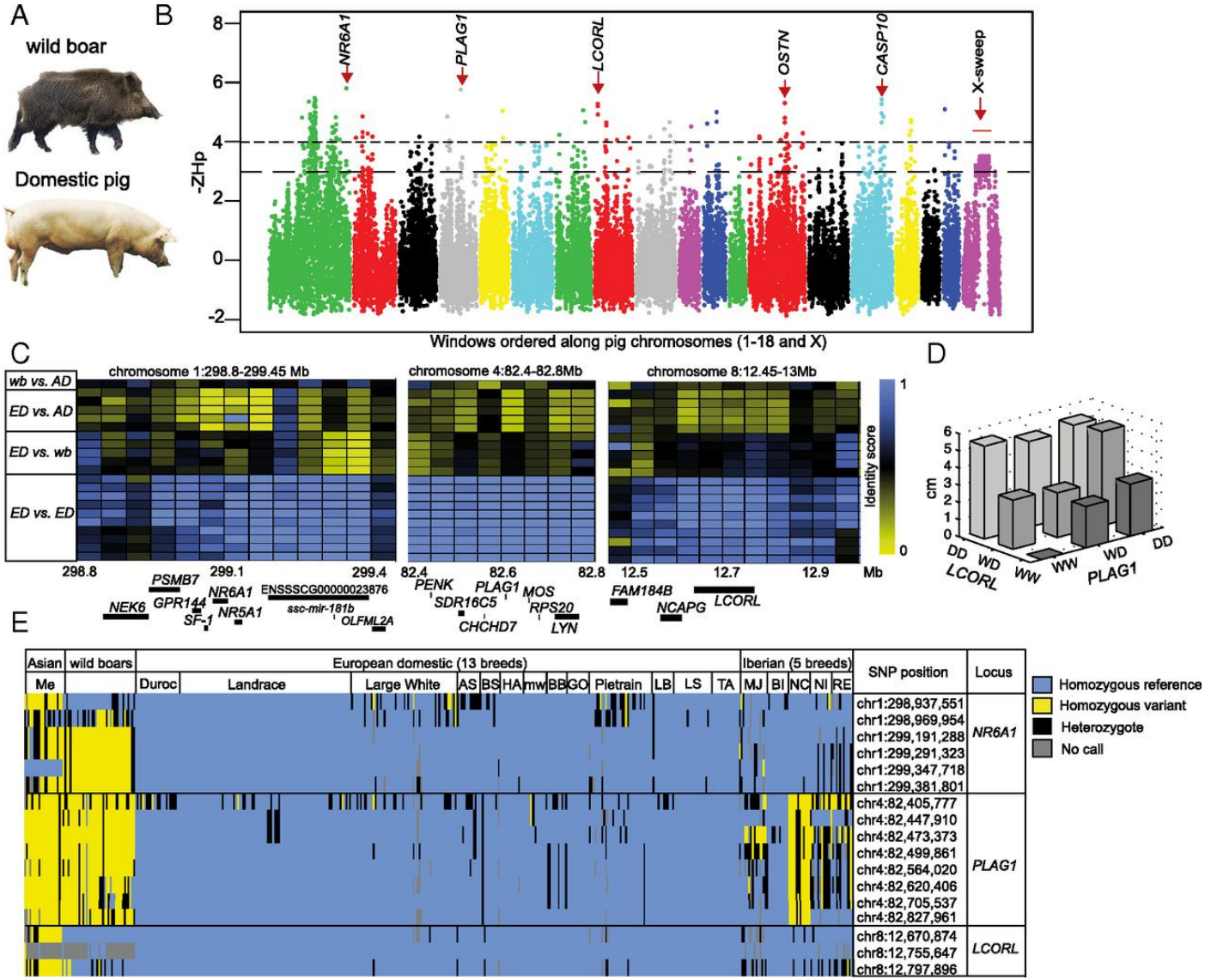
†An improved ncRNA annotation with 3,601 ncRNAs and structured elements is available as a separate track in Ensembl version 70 and for download from http://rth.dk/resources/rnannotator/susscr102.

N50, 50% of the genome is in fragments of this length or longer.

Table 1. Samples used for whole-genome shotgun sequencing

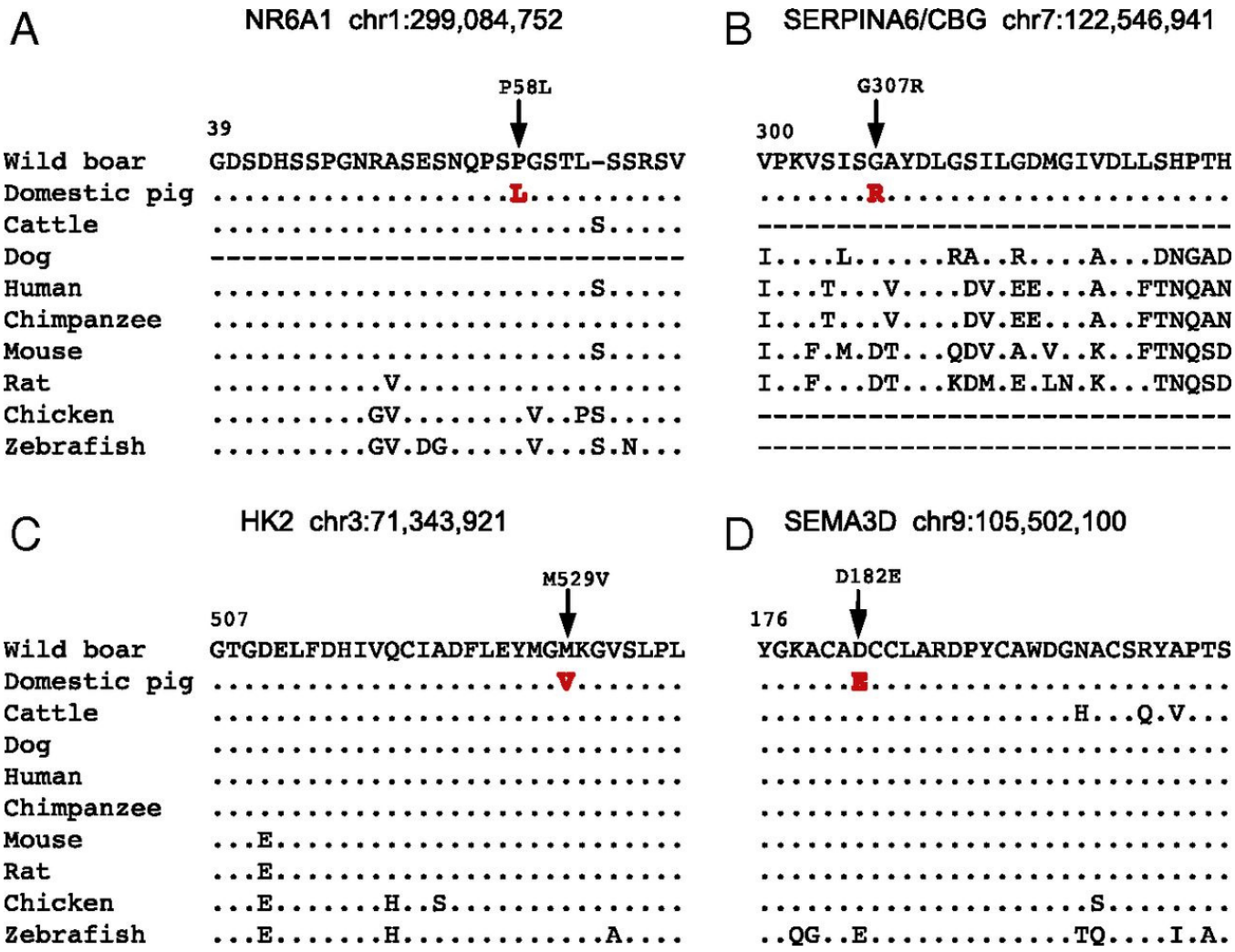
Population	Type	*n*	Sex
Dataset *i*			
Large White Uppsala	ED	8	F
Danish Landrace	ED	15	F
Danish Duroc	ED	15	F
Danish Hampshire	ED	15	F
F_2 intercross	F2	14	M/F
Large White Roslin	ED	10	F
Meishan	AD	20	F
European wild boar	EWB	20	M/F
Dataset *ii*			
Large White	ED	14	2 M/12 F
Hampshire	ED	2	2 M
Pietrain	ED	5	2 M/3 F
Duroc	ED	4	4 M
Landrace	ED	5	1 M/4 F
European wild boar	EWB	6	4 M/2 F
Asian wild boar	AWB	5	3 M/2 F
Meishan	AD	4	2 M/2 F
Xiang	AD	2	2 F
Jianquhai	AD	1	1 F
Sus scrofa (Sumatra)	OG	2	1 M/1 F
Sus barbatus	OG	1	1 M
Sus verrucosus	OG	1	1 M
Sus cebifrons	OG	1	1 F
Sus celebensis	OG	1	1 F
Phacochoerus africanus	OG	1	1 F

Dataset *i* consisted of pooled samples sequenced, SOLiD mate pair reads. Gap sizes were in the range 1,010–1,430 bp. Dataset *ii* consisted of samples sequenced individually, Illumina paired-end reads. ED, European domestic; AD, Asian domestic; EWB, European wild boar, AWB, Asian wild boar; OG, outgroup; F_2, F_2 progeny from a Large White/wild boar intercross.



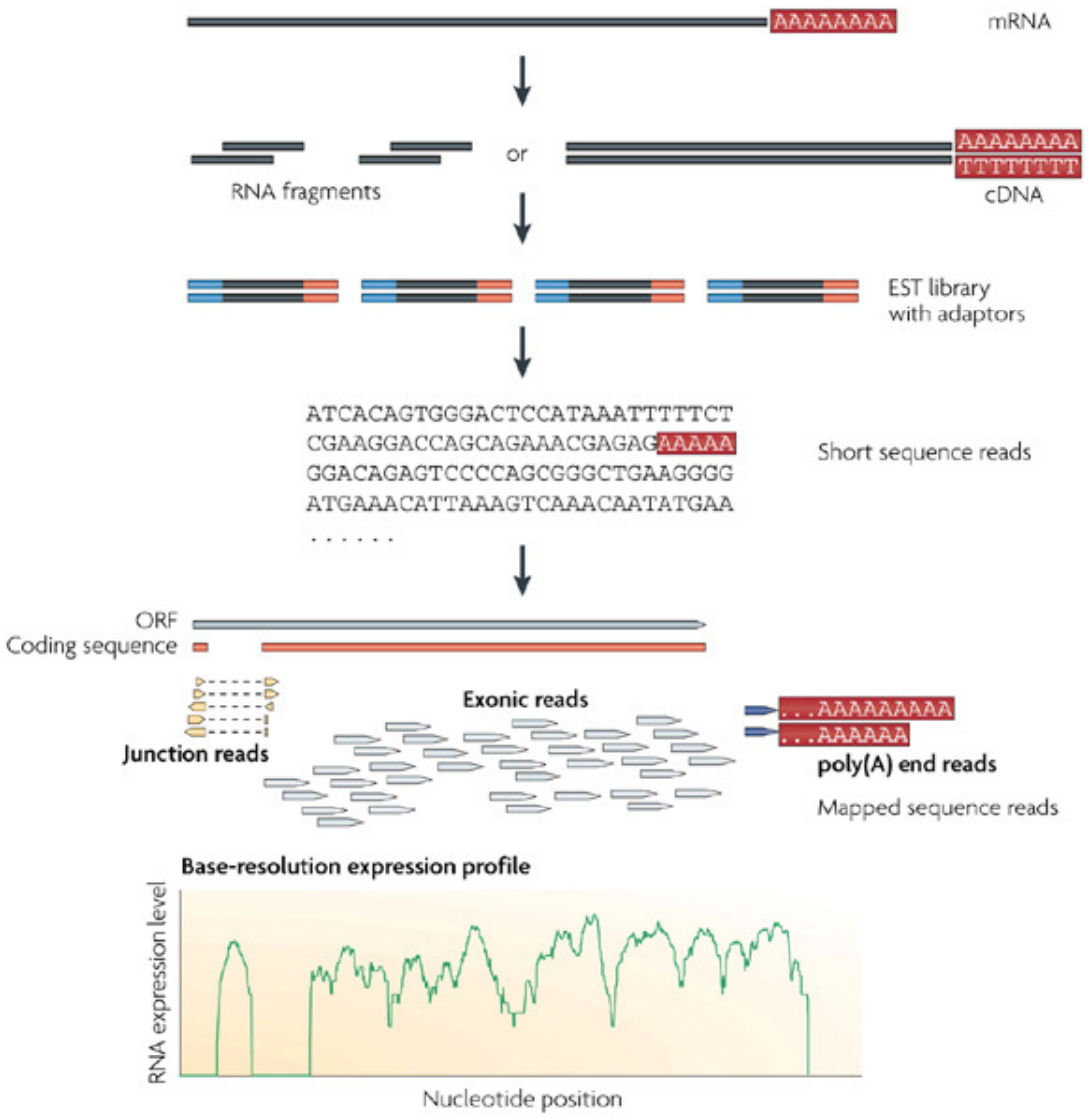
Rubin C et al. PNAS 2012;109:19529-19536

Excess of derived nonsynonymous substitutions in domestic pigs.



Rubin C et al. PNAS 2012;109:19529-19536

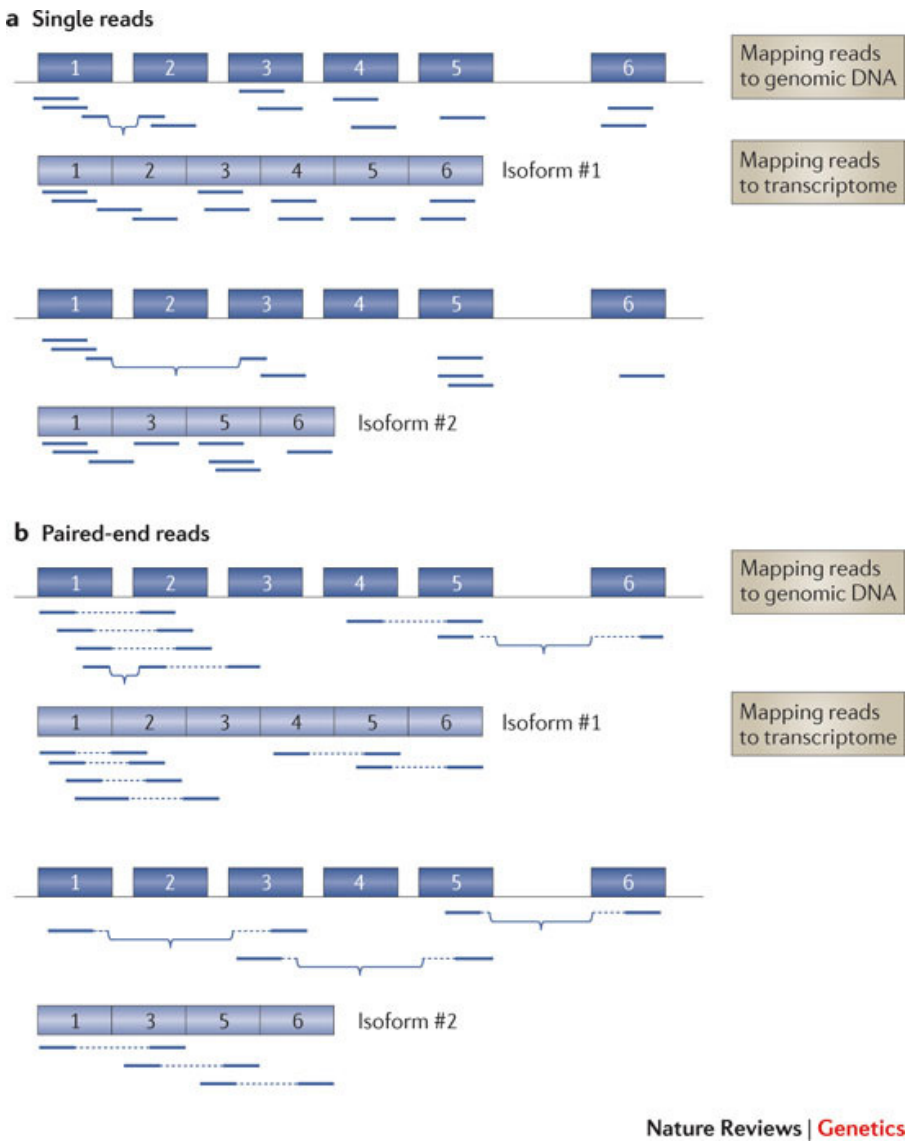
RNA-seq



Nature Reviews | Genetics

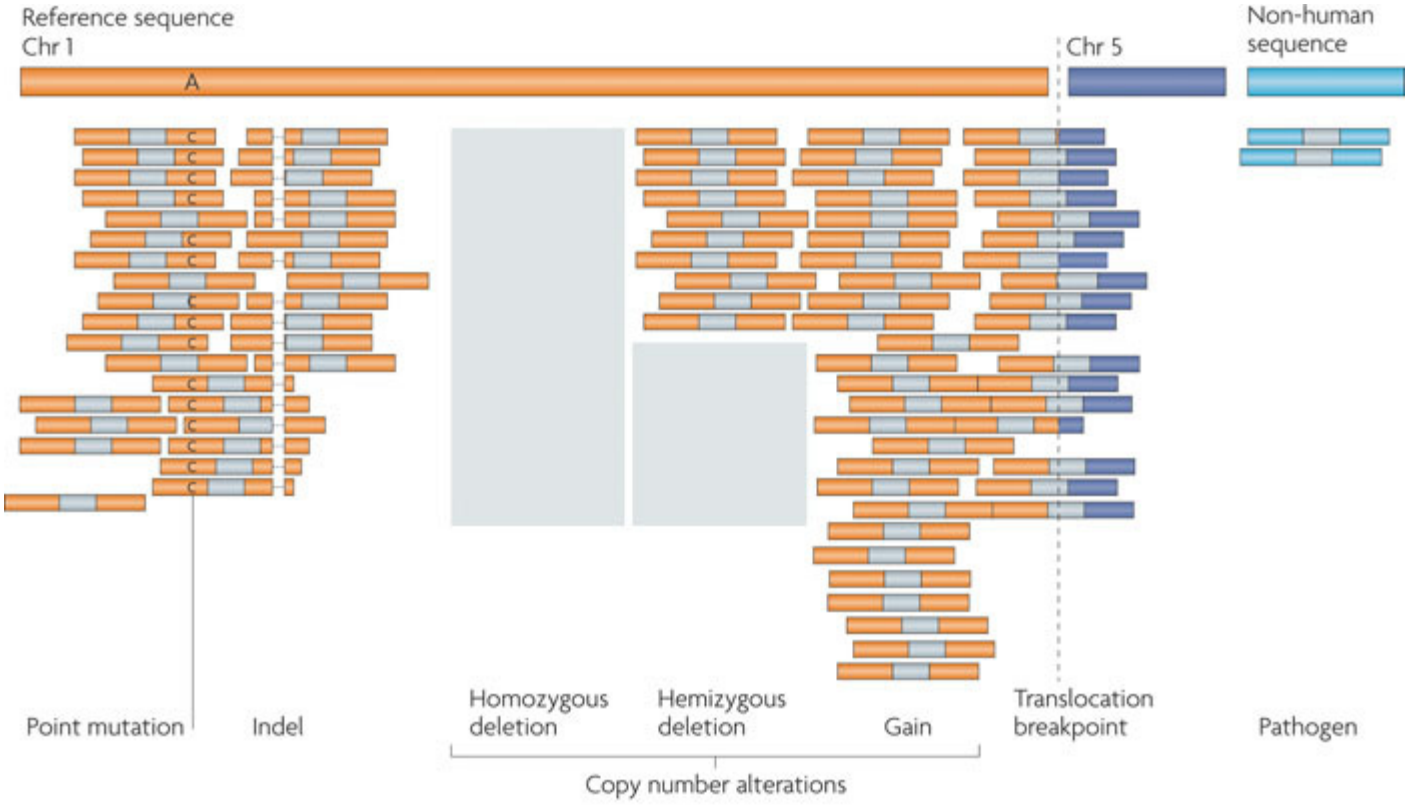
RNA-seq

RNA-seq for detection of alternative splicing events

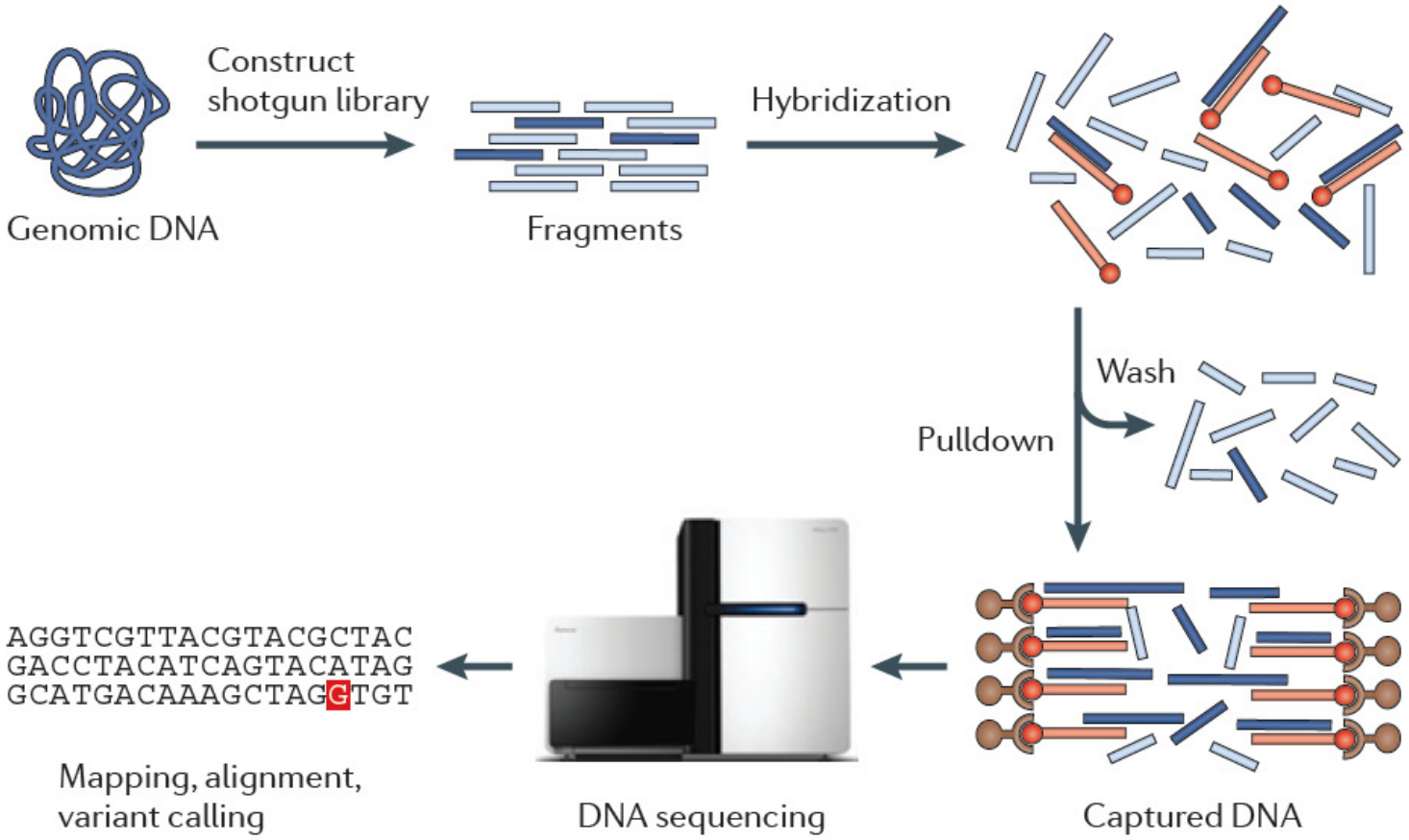


NGS in cancer genomics

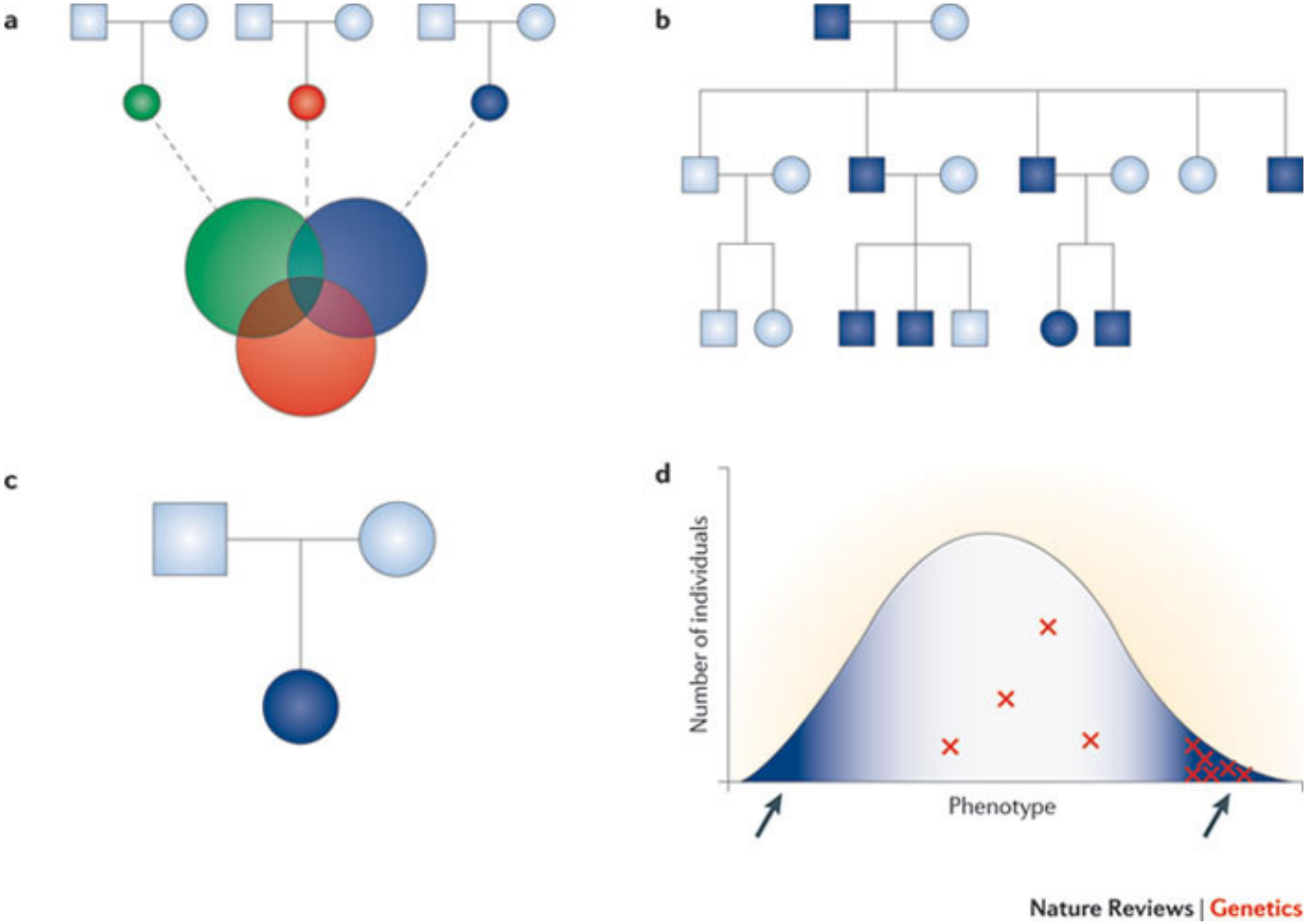
Types of genome alterations that can be detected by NGS



Exome sequencing



Strategies for finding disease-causing rare variants using exome sequencing



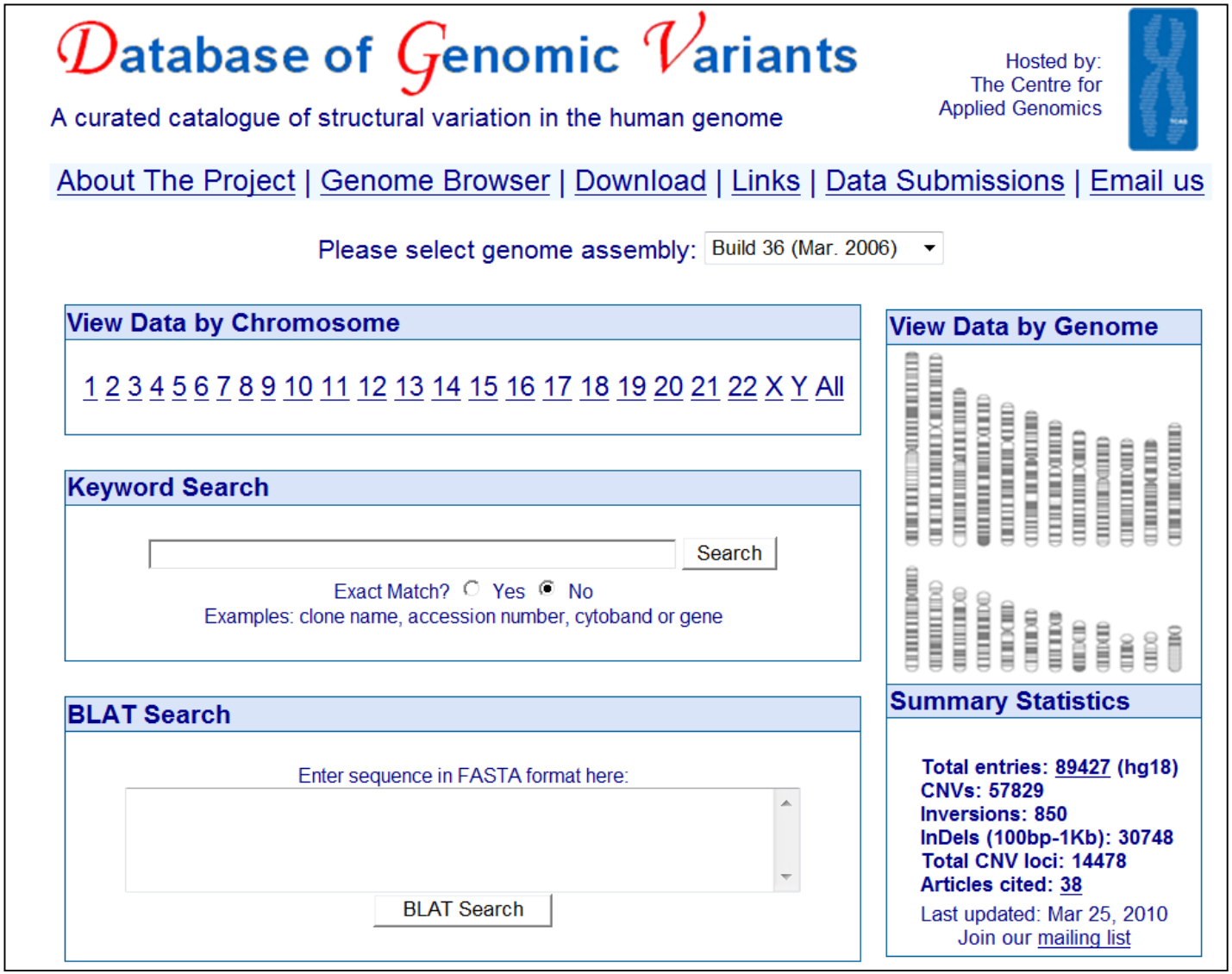
Copy Number Variation

Definitions

Copy Number Variation (CNV)

A DNA segment that is 1 kb or larger and present at variable copy number in comparison with a reference genome (Redon et al., 2006)

Intra-specific gains or losses of more than 1kb of genomic DNA (Lee et al., 2008)



http://projects.tcag.ca/variation/

Examples of CNVs determining genetic diseases in human

1) CCL3L1: HIV-1/AIDS susceptibility (Gonzalez et al., 2005)
2) FCGR3: glomerulonephritis (Aitman et al., 2006)
3) FCGR3B: systemic autoimmunity (Fanciulli et al., 2007)
4) C4: systemic lupus erythematosus (Yang et al., 2007)
5) Trypsinogen: hereditary pancreatitis (Le Marechal et al., 2006)
6) Laminin B1: autosomal dominant leukodystrophy (Padiath et al., 2006)
7) CHD7: CHARGE syndrome (Jongmans et al., 2006)
8) Alpha-synuclein: Parkinson's disease (Singleton et al., 2003)
9) APP: Alzheimer disease (Rovelet-Lecrux et al., 2006)
10) beta-defensin 2: Crohn disease (Fellermann et al., 2006)
11) Autism (e.g.: Sebat et al., 2007)
12) Schizophrenia (e.g.: Stefansson et al., 2008; Xu et al., 2008)
13) ...

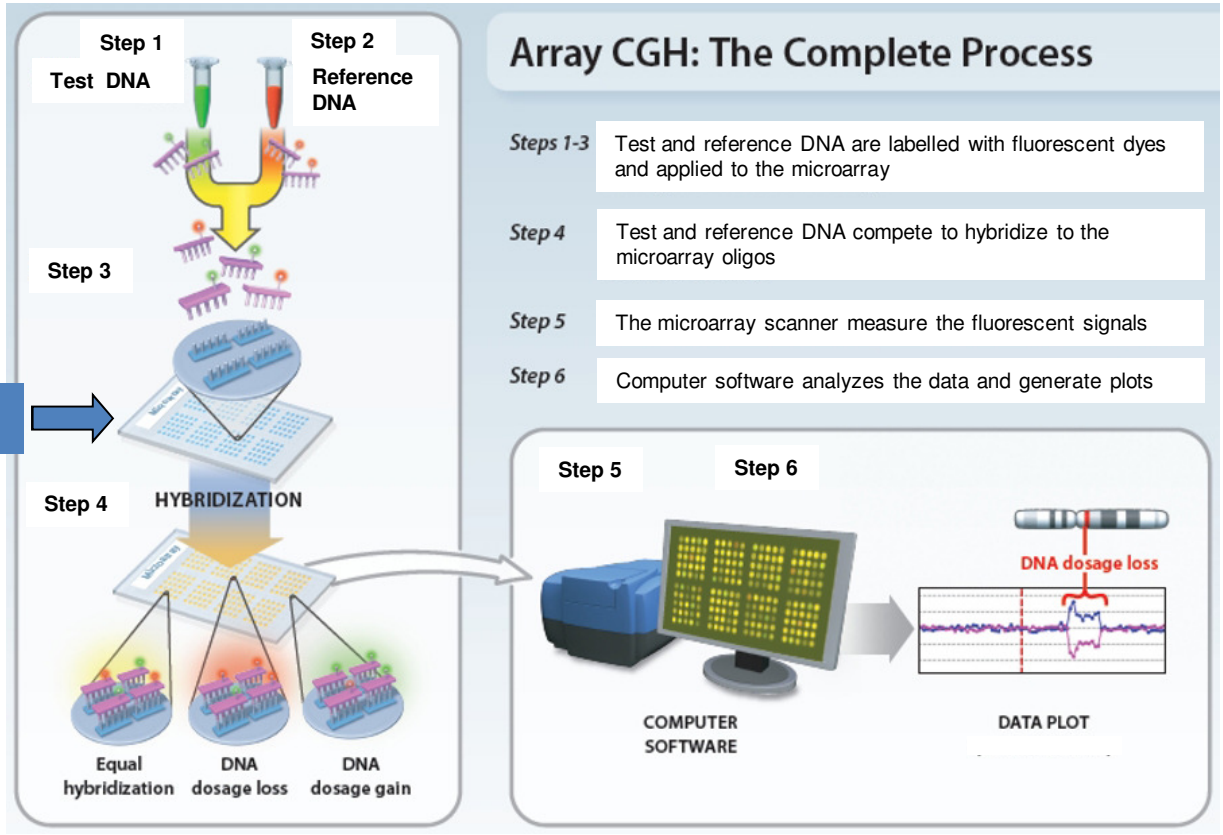
Methods to identify CNVs

High density SNP chips

Array Comparative Genome Hybridization (aCGH)

Array Comparative Genome Hybridization (aCGH)

Microarray design
(385,000 70mers)



(mod. from Theisen 2008)