Sciencexpress Report

Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays

Radoje Drmanac,[1]* Andrew B. Sparks,[1†] Matthew J. Callow,[1†] Aaron L. Halpern,[1†] Norman L. Burns,[1†] Bahram G. Kermani,[1†] Paolo Carnevali,[1†] Igor Nazarenko,[1†] Geoffrey B. Nilsen,[1†] George Yeung,[1†] Fredrik Dahl,[1†‡] Andres Fernandez,[1†] Bryan Staker,[1†] Krishna P. Pant,[1†] Jonathan Baccash,[1] Adam P. Borcherding,[1] Anushka Brownley,[1] Ryan Cedeno,[1] Linsu Chen,[1] Dan Chernikoff,[1] Alex Cheung,[1] Razvan Chirita,[1] Benjamin Curson,[1] Jessica C. Ebert,[1] Coleen R. Hacker,[1] Robert Hartlage,[1] Brian Hauser,[1] Steve Huang,[1] Yuan Jiang,[1] Vitali Karpinchyk,[1] Mark Koenig,[1] Calvin Kong,[1] Tom Landers,[1] Catherine Le,[1] Jia Liu,[1] Celeste E. McBride,[1] Matt Morenzoni,[1] Robert E. Morey,[1§] Karl Mutch,[1] Helena Perazich,[1] Kimberly Perry,[1] Brock A. Peters,[1] Joe Peterson,[1] Charit L. Pethiyagoda,[1] Kaliprasad Pothuraju,[1] Claudia Richter,[1] Abraham M. Rosenbaum,[2] Shaunak Roy,[1] Jay Shafto,[1] Uladzislau Sharanhovich,[1] Karen W. Shannon,[1‖] Conrad G. Sheppy,[1] Michel Sun,[1] Joseph V. Thakuria,[2] Anne Tran,[1] Dylan Vu,[1] Alexander Wait Zaranek,[2] Xiaodi Wu,[3] Snezana Drmanac,[1] Arnold R. Oliphant,[1] William C. Banyai,[1] Bruce Martin,[1] Dennis G. Ballinger,[1]* George M. Church,[2] Clifford A. Reid[1]

[1]Complete Genomics, Inc., 2071 Stierlin Court, Mountain View, CA 94043, USA. [2]Department of Genetics, Harvard Medical School, Cambridge, MA, USA. [3]School of Medicine, Washington University, St. Louis, St. Louis, MO, USA.

*To whom correspondence should be addressed. E-mail: rdrmanac@completegenomics.com (R.D.); dballinger@completegenomics.com (D.G.B.)

†These authors contributed equally to this work.

‡Present address: Ion Torrent Systems, San Francisco, CA, USA.

§Present address: San Diego State University, San Diego, CA, USA.

‖Present address: Life Technologies, Carlsbad, CA, USA.

Genome sequencing of large numbers of individuals promises to advance the understanding, treatment, and prevention of human diseases, among other applications. We describe a genome sequencing platform that achieves efficient imaging and low reagent consumption with combinatorial probe anchor ligation (cPAL) chemistry to independently assay each base from patterned nanoarrays of self-assembling DNA nanoballs (DNBs). We sequenced three human genomes with this platform, generating an average of 45- to 87-fold coverage per genome and identifying 3.2 to 4.5 million sequence variants per genome. Validation of one genome data set demonstrates a sequence accuracy of about 1 false variant per 100 kilobases. The high accuracy, affordable cost of $4,400 for sequencing consumables and scalability of this platform enable complete human genome sequencing for the detection of rare variants in large-scale genetic studies.

Genotyping technologies have enabled the routine assessment of common genetic variants at up to a million sites across the genome in thousands of individuals (*1*) and have increased our understanding of human genetic diversity and its biological and medical impact. Whole-genome sequencing costs have dropped from the >$100 M cost of the first human genomes (*2*, *3*) to the point where individual labs have generated genome sequences in a matter of months for material costs of as low as $48k (*4*–*12*) (table S5). Sequencing technologies, which use a variety of genomic microarray construction methodologies and sequencing chemistries (*13*–*32*), can determine human genetic diversity over an entire genome and identify common as well as rare SNPs, insertions and deletions. Despite these advances, improvements are still needed to enable the cost-effective characterization of the many hundreds of genomes required for complex disease genetic studies and for personalized disease prevention, prognosis and treatment.

We generated sequencing substrates (Fig. 1A and SOM) by means of genomic DNA fragmentation and recursive cutting with type IIS restriction enzymes and directional adaptor insertion (Fig. 1B and fig. S1). The resulting circles were then replicated with *Phi*29 polymerase (RCR) (*34*). Using a controlled, synchronized synthesis we obtained

hundreds of tandem copies of the sequencing substrate in palindrome-promoted coils of ssDNA, referred to as DNA nanoballs (DNBs) (Fig. 1C). DNBs were adsorbed onto photolithographically etched, surface modified (SOM) 25 x 75 mm silicon substrates with grid-patterned arrays of ~300nm spots for DNB binding (Fig. 1C). The use of patterned arrays increased DNA content per array and image information density relative to random genomic DNA arrays (*6*, *9*, *11*, *14*, 28). High-accuracy cPAL sequencing chemistry was then used to independently read up to 10 bases adjacent to each of eight anchor sites (Fig. 1D), resulting in a total of 31- to 35-base mate-paired reads (62 to 70 bases per DNB). cPAL is based on unchained hybridization and ligation technology (*15*, *27*, *28*, *31*), previously used to read 6-7 bases from each of four adaptor sites (26 total bases) (*28*), here extended using degenerate anchors to read up to 10 bases adjacent to each of the eight inserted adaptor sites (Fig. 1D, right) with similar accuracy at all read positions (fig. S3). This increased read length is essential for human genome sequencing.

Cell lines derived from two individuals previously characterized by the HapMap project (*33*), a Caucasian male of European decent (NA07022) and a Yoruban female (NA19240) were sequenced. NA12940 was selected to allow for a comparison of our sequence to the sequence of the same genome currently being assembled by 1000 genome project. In addition, lymphoblast DNA from a Personal Genome Project Caucasian male sample, PGP1 (NA20431) was sequenced because substantial data are available for biological comparisons (*35*–*37*). Automated cluster analysis of the four-dimensional intensity data produced raw base reads and associated raw base scores (SOM).

We mapped these sequence reads to the human genome reference assembly with a custom alignment algorithm that accommodates our read structure (fig. S4, SOM), resulting in between 124 and 241 Gb mapped and an overall genome coverage of 45- to 87-fold per genome.

To assess representational biases during circle construction we assayed genomic DNA and intermediate steps in the library construction process by quantitative PCR (QPCR) (fig. S2, SOM). This and mapped coverage showed a substantial deviation from Poisson expectation with excesses of both high and low coverage regions (fig. S5) but only a few percent of bases have coverage insufficient for assembly (Table 1). Much of this coverage bias is accounted for by local GC content in NA07022, a bias that was significantly reduced by improved adapter ligation and PCR conditions in NA19240 (fig. S5, SOM); the fraction of the genome with less than 15-fold coverage was accordingly reduced from 11% in NA07022 to 6.4% in NA19240 despite the latter having 25% less total coverage (Table 1).

Discordance with respect to the reference genome in uniquely mapping reads from NA07022 was 2.1% (range 1.4% – 3.3% per slide). However, considering only the highest scoring 85% of base calls reduced the raw read discordance to 0.47%, including about 0.1% of true variant positions.

Mapped reads were assembled into a best-fit, diploid sequence with a custom software suite employing both Bayesian and de Bruijn graph techniques (SOM). This process yielded diploid reference, variant or no-calls at each genomic location with associated variant quality scores. Confident diploid calls were made for 86 to 95% of the reference genome (Table 1), approaching the 98% that can be reconstructed in simulations. The 2% that is not reconstructed in simulations is composed of repeats that are longer than the ~400 base inserts used here and of high enough identity to prevent attribution of mappings to specific repeat copies. Longer mate-pair inserts minimize this limitation (*6*, *9*). Similar limitations affect other short read technologies.

We identified a range of 2.91 to 4.04 million SNPs with respect to the reference genome, 81 to 90% of which are reported in dbSNP, as well as short indels and block substitutions (Table 1 and table S6). Because of the use of local de novo assembly, indels were detected in sizes ranging up to 50 bp. As expected, indels in coding regions tend to occur in multiples of length 3, indicating the possible selection of minimally impacting variants in coding regions (fig. S6).

As an initial test of sequence accuracy, we compared our called SNPs with the HapMap phase I/II SNP genotypes reported for NA07022 (*1*). We fully called 94% of these positions with an overall concordance of 99.15% (Table 2) (the remaining 6% of positions were either half-called or not called). Furthermore; we fully called 96% of the Infinium (Illumina, San Diego, CA) subset of the HapMap SNPs with an overall concordance rate of 99.88%, reflecting the higher reported accuracy of these genotypes (*33*). Similar concordance rates with available SNP genotypes were observed in NA19240 (with a call rate of over 98%) and NA20431 (table S7).

We further characterized 134 of the 168 calls that were discordant with Infinium loci and Sanger sequencing of PCR products in NA07022, demonstrating that 55% of these discordances are errors in the reported HapMap genotypes (Table 2). The relationship between detection rate and read depth for about 1M Infinium HD SNPs that we subsequently genotyped in NA07022 shows that coverage of 25-fold at a position is sufficient to detect 91% of SNPs at heterozygous loci and 99% of SNPs at homozygous loci (fig. S5). Because the whole-genome false positive rate cannot be accurately estimated from known SNP loci, we tested a random subset of novel non-synonymous variants in

NA07022, a category that is enriched for errors (*10*). We extrapolated error rates from the targeted sequencing of 291 such loci, and estimated the false positive rate at about one variant per 100 kb, including <6.1 substitution-, <3.0 short deletion-, <3.9 short insertion- and <3.1 block-variants per Mb (Table 3 and table S8).

Aberrant mate-pair gaps may indicate the presence of length-altering structural variants and rearrangements with respect to the reference genome. A total of 2,126 clusters of such anomalous mate-pairs were identified in NA07022. We performed PCR-based confirmation of one such heterozygous 1,500-base deletion (fig. S7). More than half of the clusters are consistent in size with the addition or deletion of a single *Alu* repeat element.

Some applications of complete genome sequencing may benefit from maximal discovery rates, even at the cost of additional false-positives, while for others, a lower discovery rate and lower false-positive rate may be preferable. We used the variant quality score to tune call rate and accuracy (fig. S8). Additionally, novelty rate (relative to dbSNP) is also a function of variant quality score (fig. S9).

We processed the NA07022 data with Trait-o-Matic automated annotation software [as in (*12*)] yielding 1,159 annotated variants, 14 of which may have disease implications (table S10).

Because, the DNB sequencing substrates are produced by rolling-circle replication (*34*) in a uniform-temperature, solution-phase reaction with high template concentrations (> 20 billion per ml) this system avoids significant selection bottlenecks and non-clonal DNBs. This circumvents the stochastic inefficiencies of approaches that require precise titration of template concentrations for in situ clonal amplification in emulsion (*9, 14, 29*) or bridge PCR (*6, 19*).

Our patterned arrays include high-occupancy and high-density nanoarrays self-assembled on photolithography-patterned, solid-phase substrates through electrostatic adsorption of solution-phase DNBs and yield a high proportion of informative pixels (site occupancies >95%) (fig. S12A)compared to random-position DNA arrays. This results in several hundred reaction sites in the compact (~300 nm diameter) DNB produce bright signals useful for rapid imaging of the sequences (SOM). Such small DNBs also allow for high density arrays. The data set reported herein was generated with arrays with ~350 million spots at a pitch of 1.29 µm. Such a spot density and higher ones achieved in proof of concept experiments (fig. S12B), result in high image efficiency and reduced reagent consumption that enable high sequencing throughput per instrument critical for high scale human genome sequencing for research and clinical applications.

Both sequencing by synthesis (SBS) and sequencing by ligation (SBL) use chained reads, wherein the substrate for cycle N+1 is dependent on the product of cycle N; consequently errors may accumulate over multiple cycles and data quality may be affected by errors (especially incomplete extensions) occurring in previous cycles. Thus, reactions need to be driven to near completion with high concentrations of expensive high purity labeled substrate molecules and enzymes. The independent, unchained nature of cPAL avoids error accumulation and tolerates low quality bases in otherwise high quality reads, thereby decreasing reagent costs. The average sequencing consumables cost for these three genomes was under $4,400 (table S5). The raw base and variant call accuracy achieved compares favorably with other reported human genome sequences (*2–12*).

As the sequencing substrates are produced by a DNA engineering process based on modified nick-translation for directional adaptor insertion (SOM), we obtained over 90% yield in adaptor ligation; and low chimeric rates of about 1% (SOM). DNA molecules with an inserted adaptor are further enriched with PCR (SOM). This recursive process can be implemented in batches of 96 samples and extended by inserting additional adaptors to read 120 bases or more per DNB (fig. S10). The current read length is comparable to other massively parallel sequencing technologies (*6–12*).

The sequence data reported here achieve sufficient quality and accuracy for complete genome association studies, the identification of potentially rare variants associated with disease or therapeutic treatments, and the identification of somatic mutations. The low cost of consumables and efficient imaging may enable studies of several hundreds of individuals. The higher accuracy and completeness required for clinical diagnostic applications provides incentive for continued improvement of this and other technologies.

References and Notes

1. T. A. Manolio, L. D. Brooks, F. S. Collins, *J. Clin. Invest.* **118(5)** 1590 (2008).
2. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
3. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
4. S. Levy *et al.*, *PLoS Biol.* **5**, e254 (2007).
5. D. A. Wheeler *et al.*, *Nature* **452**, 872(2008).
6. D. R. Bentley *et al.*, *Nature* **456**, 53(2008).
7. J. Wang *et al.*, *Nature* **456**, 60 (2008).
8. S. M. Ahn *et al.*, *Genome Res.* **19**, 1622 (2009).
9. K. J. McKernan *et al.*, *Genome Res.* **19**, 1527 (2009).
10. T. J. Ley *et al.*, *Nature* **456**, 66 (2008).
11. D. Pushkarev, N. F. Neff, S. R. Quake, *Nat. Biotechnol.* **27**, 847 (2009).
12. J. I. Kim *et al.*, *Nature*. **460**, 1011 (2009).
13. R. Drmanac *et al.*, *Genomics* **4**(2)**,** 114 (1989).
14. R. Drmanac, R. Crkvenjakov, *Scientia Yugoslavica*, **16** (1-2), 97 (1990).
15. R. Drmanac *et al.*, *Science* **260**, 1649 (1993).

16. P.C. Cheesman, US patent 5,302,509 (1994).
17. R. Drmanac, World Intellectual Property Organization WO/1995/009248 (1995).
18. M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlen, P. Nyren, *Anal. Biochem.* **242**, 84 (1996).
19. C. P. Adams, S. J. Kron, U.S. Patent 5,641,658 (1997).
20. P. M. Lizardi *et al.*, *Nat. Genet.* **19**, 225 (1998).
21. S. C. Macevicz, U.S. Patent 5,750,341 (1998).
22. S. Drmanac, D Kita, *et al.*, *Nat. Biotechnol.* **16,** 54 (1998).
23. R. D. Mitra, G. M. Church, *Nucleic Acids Res.* **27**, e34 (1999).
24. S. Brenner *et al.*, *Nat. Biotechnol.* **18**, 630 (2000).
25. I. Braslavsky, B. Hebert, E. Kartalov, S. R. Quake, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3960 (2003).
26. R. D. Mitra, J. Shendure, J. Olejnik, O. E. Krzymanska, G. M. Church, *Anal. Biochem.* **320**, 55 (2003).
27. R. Drmanac *et al.*, World Intellectual Property Organization WO/2004/076683 (2004).
28. J. Shendure *et al.*, *Science* **309,** 1728 (2005).
29. M. Margulies *et al.*, *Nature* **437**, 376 (2005).
30. T.D. Harris *et al.*, *Science* **320**, 106 (2008).
31. A. Pihlak *et al.*, *Nat Biotechnol.* **26(6)**, 676 (2008).
32. J. Shendure, H. Ji, *Nat Biotechnol.* **26,** 1135 (2008).
33. The International HapMap Consortium, *Nature* **449**, 851 (2007).
34. L. Blanco *et al.*, *J Biol Chem.* **264**, 8935 (1989).
35. K Zhang *et al*., *Nature Methods* **6(8),** 613 (2009).
36. M.P. Ball, *et al*., *Nature Biotechnol.* **27,** 361 (2009).
37. J. B. Li *et al*., *Genome Research* Jul 13. PMID: 19525355 (2009).
38. We acknowledge and the ongoing contributions and support of all Complete Genomics employees and R. Mercado for manuscript preparation. Some of this work was supported from PersonalGenomes.org, and NHLBI. Data has been deposited at NCBI: reads in the Short Read Archive (SRA), accession SRA008092; and SNPs in dbSNP, accessions ss161884913 to ss175323894. Employees of Complete Genomics have stock options in the company and DGB has stock in Perlegen Sciences. Complete genomics has filed several patents on this work.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1181498/DC1
Materials and Methods
Figs. S1 to S12
Tables S1 to S9
References

3 September 2009; accepted 23 October 2009
Published online 5 November 2009;
10.1126/science.1181498
Include this information when citing this paper.

Fig. 1. Amplified DNA nanoarray platform. (**A**) Schematic flow diagram of the process used. (**B**) Library construction schematic (SOM, fig. S1). (**C**) DNB production (fig. S11, SOM) and nanoarray formation (SOM) schematics. (**D**) Schematic of combinatial probe anchor ligation (cPAL) products (SOM).

Table 1. Summary information from mapping and assembly of three genomes. All variations are with respect to the National Center for Biotechnology Information (NCBI) version 36 human genome reference assembly. Novel variations were ascertained by comparison to dbSNP (JDW, release 126; NA18507 (*6*), release 128; all other genomes, release 129). NA18507 and NA19240 are Yoruban HapMap samples, which may explain the number of SNPs and novelty rates. In partially called regions of the genome, one allele could be called confidently but not the other. The high call rate in NA19240 reflects reduced library bias (fig. S5).

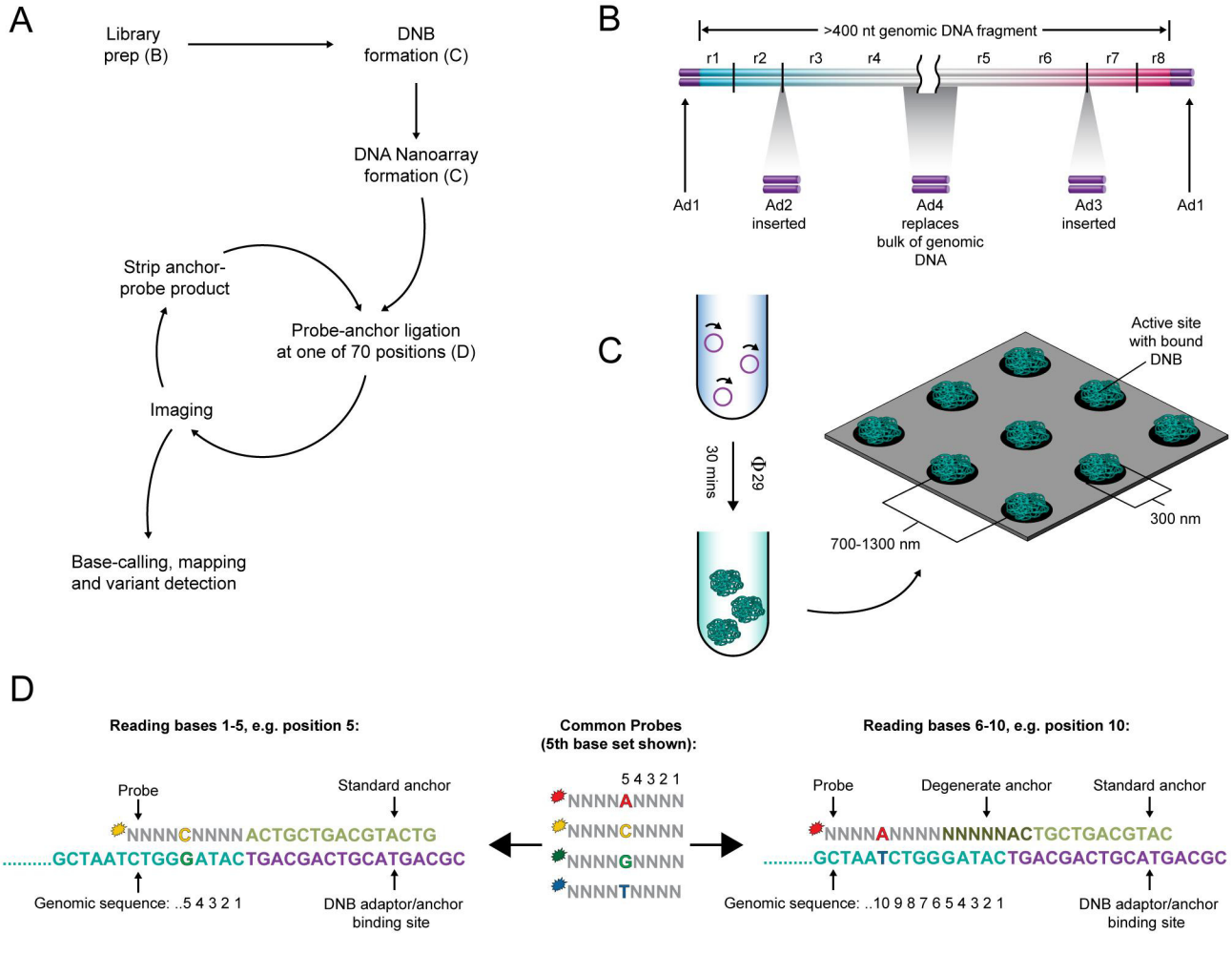
Sample	Mapped sequence (Gb)	Average Coverage depth (fold)	Percent of genome called		SNPs		Indels		Insertion:Deletion ratio
			Fully	Partially	Total	Novel %	Total	Novel %	
Genomes sequenced by Complete Genomics:									
NA07022 (*35*)	241	87	91%	2%	3,076,757	10%	337,604	37%	1.0
NA19240 (*36*)	178	63	95%	1%	4,042,801	19%	496,149	42%	0.96
NA20431 (*37*)	124	45	86%	3%	2,905,517	10%	269,794	37%	1.0
Genomes previously published:									
NA18507 (*6*)	135	47	–	–	4,139,196	26%	404,416	50%	0.77
NA18507 (*9*)	49	31	–	–	3,866,085	19%	226,529	33%	0.72
JCV (*3*)	21	7	–	–	3,213,401	15%	851,575	–	–
JDW (*4*)	21	7	–	–	3,322,093	18%	222,718	51%	0.4

Table 2. Concordance with genotypes for NA07022 generated by the HapMap Project (release 24) and the highest quality Infinium assay subset of those genotypes, as well as genotyping on Illumina Infinium 1M assay. Discordances with reported HapMap Infinium genotypes were verified by Sanger sequencing (SOM).

			Infinium 1M	HapMap phase I&II SNPs	HapMap Infinium subset	HapMap Infinium SNPs tested for accuracy by Sanger sequencing		
Published Concordance			–	99.03%	99.94%			
NA07022	# reported		1 M	3.9 M	143 K	These data correct	These data incorrect	% affirmed
	% called		95.98%	94. 39%	96.00%			
	% locus concordance		99.89%	99.15%	99.88%			
	HapMap genotype calls	Homozygous ref	99.96%	99.34%	99.96%	18	2	90%
		Heterozygous	99.78%	99.39%	99.80%	28	46	38%
		Homozygous alt	99.81%	98.14%	99.84%	28	12	70%

Table 3. False positive rates and FDRs were calculated for the entire set of variations called in NA07022 by extrapolating the heterozygous (Het) FDRs calculated from comparative Sanger sequencing of 291 selected novel variants (table S8) to all variants. This is a conservative approach (detailed in SOM text). The total number of all types of false positive variants is estimated at 7.5-16.1 per Mb.

Variation Type	Total detected	Novel	Het novel FDR (table S8)	Estimated false positives on genome	Estimated false positives / Mbp	Estimated FDR
SNP	3,076,869	310,690	2-6%	7k-17k	2.3-6.1	0.2-0.6%
Deletion	168,726	61,960	8-14%	5k-8k	1.8-3.0	3.0-5.0%
Insertion	168,909	61,933	11-18%	7k-11k	2.3-3.9	3.9-6.5%
Block substitution	62,783	30,445	11-29%	3k-9k	1.1-3.1	5.2-13.9%


A
Library prep (B)
DNB formation (C)
DNA Nanoarray formation (C)
Probe-anchor ligation at one of 70 positions (D)
Imaging
Strip anchor-probe product
Base-calling, mapping and variant detection
B
>400 nt genomic DNA fragment
r1
r2
r3
r4
r5
r6
r7
r8
Ad1
Ad2 inserted
Ad4 replaces bulk of genomic DNA
Ad3 inserted
Ad1
C
30 mins
Φ29
Active site with bound DNB
300 nm
700-1300 nm
D
Reading bases 1-5, e.g. position 5:
Probe
Standard anchor
NNNNCNNNN ACTGCTGACGTACTG
.....GCTAATCTGGGATACTGACGACTGCATGACGC
Genomic sequence: ..5 4 3 2 1
DNB adaptor/anchor binding site
Common Probes (5th base set shown):
5 4 3 2 1
NNNNANNNN
NNNNCNNNN
NNNNGNNNN
NNNNTNNNN
Reading bases 6-10, e.g. position 10:
Probe
Degenerate anchor
Standard anchor
NNNNANNNNNNNNNACTGCTGACGTAC
.....GCTAATCTGGGATACTGACGACTGCATGACGC
Genomic sequence: ..10 9 8 7 6 5 4 3 2 1
DNB adaptor/anchor binding site