SUPPLEMENTARY MATERIAL

TITLE

Megabase-scale haplotypes of germline and cancer genomes using linked-read sequencing

Authors:

Grace X.Y. Zheng[1,†], Billy T. Lau[2,†], Michael Schnall-Levin[1], Mirna Jarosz[1], John M. Bell[2], Christopher M. Hindson[1], Sofia Kyriazopoulou-Panagiotopoulou[1], Donald A. Masquelier[1], Landon Merrill[1], Jessica M. Terry[1], Patrice A. Mudivarti[1], Paul W. Wyatt[1], Rajiv Bharadwaj[1], Anthony J. Makarewicz[1], Yuan Li[1], Phillip Belgrader[1], Andrew D. Price[1], Adam J. Lowe[1], Patrick Marks[1], Gerard M. Vurens[1], Paul Hardenbol[1], Luz Montesclaros[1], Melissa Luo[1], Lawrence Greenfield[1], Alexander Wong[1], David E. Birch[1], Steven W. Short[1], Keith P. Bjornson[1], Pranav Patel[1], Erik S. Hopmans[2], Christina Wood[3], Sukhvinder Kaur[1], Glenn K. Lockwood[1], David Stafford[1], Joshua P. Delaney[1], Indira Wu[1], Heather S. Ordonez[1], Susan M. Grimes[2], Stephanie Greer[3], Josephine Y. Lee[1], Kamila Belhocine[1], Kristina M. Giorda[1], William H. Heaton[1], Geoffrey P. McDermott[1], Zachary W. Bent[1], Francesca Meschi[1], Nikola O. Kondov[1], Ryan Wilson[1], Jorge A. Bernate[1], Shawn Gauby[1], Alex Kindwall[1], Clara Bermejo[1], Adrian N. Fehr[1], Adrian Chan[1], Serge Saxonov[1], Kevin D. Ness[1], Benjamin J. Hindson[1], Hanlee P. Ji[2,3]

† These authors contributed equally to this work.

Institutions:

[1]10X Genomics, Pleasanton CA, United States, 94305

[2]Stanford Genome Technology Center, Stanford University, Palo Alto, CA, United States, 94304

[3]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States, 94305

Corresponding Authors:

Hanlee P. Ji

Email: genomics_ji@stanford.edu

Benjamin J. Hindson

Email: ben@10xgenomics.com

Outline of Supplementary Material

Additional Supplementary Tables as separate files.

Supplementary Table 1. Coverage statistics of all samples with WGS data. Patient 1532 was sequenced using both standard short read WGS and barcode libraries from the 10X GemCode platform.

Samples	Phased					Illumina TruSeq	Patient 1532			
	NA12878	NA12877	NA12882	NA20847	NCI-H2228	NA12878	Illumina TruSeq		Phased	
							Normal	Malignant	Normal	Malignant
Mapped sequence (Gb)	101.09	88.98	93.26	83.65	82.21	88.27	127.7	149.9	80.93	83.26
Average haploid coverage	37	34	36	32	32	34	44.6	52.3	31	32
Coverage of genome (%)	90.11%	90.16%	90.15%	89.90%	89.70%	90.31%	99.26%	99.26%	89.85%	89.53%
10X or greater sequence coverage (%)	86.20%	86.22%	86.31%	84.30%	80.33%	90.13%	99.00%	99.03%	81.30%	78.15%
20X or greater sequence coverage (%)	72.68%	70.22%	72.05%	67.80%	60.69%	87.59%	98.27%	98.50%	60.02%	56.85%
30X or greater sequence coverage (%)	55.57%	51.46%	54.80%	47.70%	44.01%	73.47%	93.47%	94.94%	40.12%	39.10%

Supplementary Table 2. Shearing size of phased samples. Samples subject to WGS and exome analysis are sheared to the sizes listed in the table before end-repair and ligation of P7 adaptor.

Phased Sample Shearing Sizes	
Sample	**Size of the library (bp)**
NA12878 WGS	250
NA12877 WGS	500
NA12882 WGS	500
NA20847 WGS	250
Patient samples	250
All exome samples	250

Supplementary Table 3. Downsampling sequence data to determine WGS phasing performance.

NA12878						
Coverage	37	26	23	17	13	10
% SNPs phased	99	97	97	96	95	93
% genes phased (<100 kb)	97	96	96	95	93	91
N50 phase block (bases)	2,834,437	2,523,822	2,341,608	1,868,683	1,496,897	1,124,096
Longest phase block (bases)	14,557,822	12,014,277	10,312,095	11,893,726	7,224,068	6,162,317
SNV short switch error rate (%)	0.01%	0.21%	0.25%	0.34%	0.51%	0.84%
SNV long switch error rate (%)	0.01%	0.02%	0.02%	0.02%	0.02%	0.03%

Supplementary Table 4. Summary of SNVs in the nuclear trio, NA20847, and patient 1532 normal-tumor pair.

	NA12878	NA12877	NA12882	NA20847	Patient 1532 normal	Patient 1532 tumor
SNVs	3,743,419	3,805,278	3,519,124	3,267,971	3,322,293	3,292,639
Heterozygous SNVs	2,377,169	2,252,548	2,140,360	1,896,984	1,955,426	1,890,051
Indels	889,946	787,433	516,580	-	714,475	718,416
Heterozygous indels	734,685	641,090	346,153	-	447,794	448,663

Supplementary Table 5. Genomic deletions of NA12878 (a) and NA12882 (b).

(a) NA12878 deletions discussed in the text								
Predicted deletion start		Predicted deletion end						
Chr	Breakpoint 1	Chr	Breakpoint 2	Quality score	# Barcode overlap	# Paired reads	# Split reads	Read pair likelihood ratio
3	162,512,134 - 162,512,135	3	162,626,332 - 162,626,335	472	41	9	5	176
1	189,704,510 - 189,704,521	1	189,783,385 - 189,783,396	458	37	6	0	70
6	78,950,000 - 78,960,000	6	79,040,000 - 79,050,000	456	36	0	1	14
8	39,220,000 - 39,230,000	8	39,390,000 - 39,400,000	455	36	0	1	14
5	104,432,115 - 104,432,116	5	104,503,670 - 104,503,673	404	33	7	3	123
(b) NA12882 deletions discussed in the text								
Predicted deletion start		Predicted deletion end						
Chr	Breakpoint 1	Chr	Breakpoint 2	Quality score	# Barcode overlap	# Paired reads	# Split reads	Read pair likelihood ratio
5	104,432,113 - 104,432,116	5	104,503,670 - 104,503,673	633	58	10	7	214
8	39,231,935 - 39,231,952	8	39,387,240 - 39,387,257	584	50	4	0	49
3	162,512,134 - 162,512,137	3	162,626,332 - 162,626,335	370	33	9	5	175

Supplementary Table 6. Inheritance pattern of the deletion candidates and summary of reads from targeted sequencing to validate the deletion candidates.

Inheritance Table								NA12878 breakpoint (mother)			NA12877 breakpoint (father)			NA12882 breakpoint (child)			% of bases with low mappability
Chr	Location	NA12878		NA12877		NA12882		across	beyond	ambiguous	across	beyond	ambiguous	across	beyond	ambiguous	
High scoring SV candidates:																	
1	189,704,509 - 189,783,359	**1**	**2**	**3**	**4**	**2**	**3**	332	609	0	0	0	0	0	0	0	0%
3	162,512,134 - 162,626,335	**1**	**2**	**3**	**4**	**1**	**3**	671	860	0	0	0	0	539	847	0	0.22%
5	104,432,113 - 104,503,673	**1**	**2**	**NP**	**NP**	**1**	**NP**	199	350	0	0	0	0	158	323	0	0.02%
6	78,967,194 - 79,036,419	**1**	**2**	**3**	**4**	**2**	**4**	48	698	0	52	253	5	0	0	0	1.86%
8	39,232,074 - 39,387,229	**1**	**2**	**3**	**4**	**1**	**4**	346	937	0	0	0	0	417	904	0	10.21%
Low scoring SV candidates:																	
5	99,400,881 - 99,715,015	**1**	**2**	**3**	**4**	**1**	**4**	35	212	8776	0	216	7094	18	166	7155	61.65%
14	37,631,609 - 37,771,228	**1**	**2**	**3**	**4**	**2**	**4**	552	20	223	558	28	223	395	30	173	10.50%
14	106,932,640 - 107,174,931*	**1**	**2**	**3**	**4**	**2**	**4**	2065	4401	46	1777	4657	82	1879	4382	25	9.59%

underlined = deletion candidate

NP = unphased due to lack of heterozygous SNPs

"across breakpoint" indicates a soft - clipped chimeric sequence

"beyond breakpoint" indicates a non - clipped read aligning on the opposite side of the breakpoints from its associated primer - probe within 1 Kb of breakpoint

"ambiguous" indicates aligning beyond the opposite breakpoint when it is a control (i.e. oriented in wrong direction or inside breakpoints)

mappability is based on UCSC 75mer and describes 1 kb regions to the outside of both breakpoints

*** in a VDJ recombination region**

Supplementary Table 8. Comparison of structural variants called in NA12878 against deletions from Pendleton et. al., 2015.

chr	breakpoint1 start	breakpoint1 stop	chr	breakpoint2 start	breakpoint2 stop	quality score	"confident" set from Pendleton et. al.[1]	"relaxed" set from Pendleton et. al.[1]	short read callers support
4	34770000	34780000	4	34830000	34840000	930	Y	Y	
2	52749686	52749689	2	52785268	52785270	834	Y	Y	
1	72766324	72766327	1	72811837	72811840	734	Y	Y	
7	54280000	54290000	7	54370000	54380000	560			Kidd et. al.[2]
20	1540000	1550000	20	1600000	1610000	505			
3	162512134	162512135	3	162626332	162626335	472		Y	
1	189704510	189704521	1	189783385	189783396	458	Y	Y	
6	78950000	78960000	6	79040000	79050000	456	Y	Y	
8	39220000	39230000	8	39390000	39400000	455		Y	
5	104432115	104432116	5	104503670	104503673	404	Y	Y	
1	152530000	152540000	1	152590000	152600000	402	Y	Y	
16	34380000	34390000	16	34740000	34750000	402			Kidd et. al.[2]
4	161020000	161030000	4	161080000	161090000	369		Y	
2	34680000	34690000	2	34740000	34750000	363	Y	Y	
11	108585765	108585777	13	21727722	21727734	323			Pendleton et. al.[1]
15	64970000	64980000	X	7100000	7110000	281			
13	47250000	47260000	13	107250000	107260000	269			
15	85510000	85520000	3	189600000	189610000	242			
4	34770000	34780000	4	34880000	34890000	237			
7	88420000	88430000	8	9070000	9080000	218			

Supplementary Table 9. *ALK-EML4* and *ALK-PTPN3* gene fusions called in exome data from H2228 cell line.

Predicted SV start		Predicted SV end						
Chr	**Breakpoint 1**	**Chr**	**Breakpoint 2**	**Quality score**	**# Barcode overlap**	**# Paired reads**	**# Split reads**	**Read pair likelihood ratio**
2	29,435,765 - 29,452,584	2	42,471,639 - 42,484,792	128	17	0	1	14
2	42,542,089 - 42,545,770	9	102,818,990 - 102,821,171	109	13	1	0	12

Supplementary Table 10. Summary of candidate deleterious mutations in Patient 1532 normal and tumor pair.

Sequencing comparison between short read and phased results																			
Short read WGS analysis													Phasing						
														Region phased?					
Chr	Position	Ref	Alt	Genotype	Coverage	Consequence	Gene	cDNA pos	Exon	oAA	nAA	CADD Phred	Status	Score	N	N haplotype block	T	T haplotype block	T hap #
1	115,258,747	C	T	0/1	22,22	non-synonymous	*NRAS*	289	2/7	G	D	35	phased	179	Y	115,001,691 - 116,205,021	Y	115,001,691 - 115,308,500	2
1	119,575,884	C	G	0/1	43,13	non-synonymous	*WARS2*	760	6/6	E	Q	35	phased	255	Y	118,674,901 - 119,999,642	Y	118,618,567 - 119,845,956	1
2	206,165,385	C	T	0/1	28,23	stop gain	*PARD3B*	2317	17/22	R	*	36	phased	51	Y	205,671,849 - 208,192,741	Y	205,701,987 - 206,280,521	1
6	136,476,831	G	T	0/1	32,14	non-synonymous	*PDE7B*	329	4/4	D	Y	29.4	phased	223	Y	136,144,830 - 136,515,751	Y	136,340,592 - 136,515,751	1
7	104,766,776	G	A	0/1	32,21	non-synonymous	*SRPK2*	255	4/6	H	Y	25.6	phased	255	Y	103,166,875 - 105,035,389	Y	104,158,828 - 105,737,510	2
7	131,849,047	C	A	0/1	33,18	stop gain	*PLXNA4*	4583	24/32	E	*	45	phased	32	Y	130,254,925 - 133,924,029	Y	131,181,808 - 133,951,513	1
8	23,190,978	C	T	0/1	10,16	non-synonymous	*LOXL2*	1272	5/14	C	Y	28.6	phased	126	Y	22,021,130 - 28,213,699	Y	23,167,609 - 23,818,687	1
10	91,522,548	A	G	1/1	1,21	non-synonymous	*KIF20B*	5010	29/33	K	E	29.6	homozygote	n/a	Y	90,151,393 - 93,064,626	Y	90,669,755 - 93,025,598	1
12	57,908,971	T	C	0/1	39,19	non-synonymous	*MARS*	6	1/4	I	T	25.6	phased	255	N	n/a	Y	57,583,486 - 57,908,971	2
12	112,512,522	G	A	0/1	39,25	stop gain	*NAA25*	1072	9/24	R	*	39	phased	255	Y	112,492,626 - 112,515,605	Y	112,389,476 - 112,527,982	1
13	46,541,951	G	A	0/1	35,25	stop gain	*ZC3H13*	4358	15/19	R	*	45	phased	255	Y	45,740,791 - 50,237,729	Y	45,939,425 - 47,598,625	2
17	7,578,211	C	T	0/1	10,22	non-synonymous	*TP53*	638	5/7	R	Q	37	phased	140	Y	3,583,844 - 8,825,330	Y	7,533,025 - 8,220,432	2
17	10,355,371	C	T	1/1	2,26	non-synonymous	*MYH4*	3736	27/40	E	K	34	homozygote	n/a	Y	8,855,408 - 15,404,987	Y	10,222,462 - 11,635,630	1
17	61,766,931	G	A	0/1	21,17	non-synonymous	*MAP3K3*	1058	13/18	R	H	32	not phased	3	N	n/a	N	n/a	n/a
19	30,164,924	G	A	0/1	25,25	non-synonymous	*PLEKHF 1*	280	2/2	D	N	34	not phased	3	Y	29,141,047 - 30,418,862	Y	29,141,047 - 30,418,862	n/a
X	41,007,638	C	T	0/1	35,21	non-synonymous	*USP9X*	1460	11/44	A	V	32	not phased	0	N	n/a	N	n/a	n/a
X	152,027,456	T	C	0/1	46,12	non-synonymous	*NSDHL*	671	5/9	V	A	25.9	not phased	0	N	n/a	N	n/a	n/a

Supplementary Table 11. Tumor-specific structural variants in Patient 1532.

Linked-read analysis					Breakdancer analysis						Read support				
Breakpoint 1		Breakpoint 2		Q score	Breakpoint 1		Breakpoint 2		SV class	Score	Breakpoint 1		Breakpoint 2		# Reads spanning breakpoint
Chr	Position	Chr	Position		Chr	Position	Chr	Position			Chr	Position	Chr	Position	
6	72,327,710 - 72,327,713	6	72,784,361 - 72,784,364	222	6	72,327,703	6	72,784,371	deletion	88	6	72,327,710	6	72,784,364	7
8	106,033,196 - 106,033,199	8	120,871,951 - 120,871,952	220	8	106,033,426	8	120,872,188	inversion	99	8	106,033,199	8	120,871,952	18
11	108,585,765 - 108,585,779	13	21,727,797 - 21,727,811	229	11	108,585,849	13	21,750,514	translocation	99	11	108,585,748	13	21,750,678	31
16	6,530,000 - 6,540,000	16	6,610,000 - 6,620,000	221	16	6,542,212	16	6,604,868	deletion	96	16	6,542,223	16	6,604,872	38
16	33,410,000 - 33,420,000	6	370,000 - 380,000	222	16	33,428,374	6	382,330	translocation	99	16	33,428,530	6	382,461	77

Q score refers to quality.

Supplementary Table 12. List of copy number alterations determined by linked-read analysis and short read segmentation.

***** Linked-reads filter considers prediction quality score and interval size. Added validation was conducted by counting barcodes in the linked-read data**.**

Chr	Interval		Aberration type	Copy number	Normal mean barcode count (50 kb window)	Tumor mean barcode count (50kb window)	Tumor/Normal mean barcode ratio (50 kb window)	Normal mean read count (50 kb window)	Tumor mean read count (50 kb window)	Tumor/Normal mean read ratio (50 kb window)	*Predicted by linked-reads?
1	3,667,863	34,284,328	Deletion	1.4	1,246	668	0.54	6,344	3,913	0.62	N
2	213,693,504	215,882,694	Deletion	1.2	1,727	966	0.56	10,777	6,853	0.64	Y
4	152,988,593	190,539,790	Deletion	1.6	1,667	1,118	0.67	9,995	7,678	0.77	N
5	132,833,428	135,468,042	Deletion	1.4	1,383	741	0.54	7,217	4,532	0.63	Y
5	166,595,763	167,913,468	Deletion	1.3	1,632	955	0.59	9,657	6,457	0.67	Y
6	24,208,288	24,266,019	Amplification	3.3	1,710	2,610	1.53	10,273	20,816	2.03	Y
6	72,327,538	72,784,577	Deletion	1.2	1,835	1,022	0.56	11,470	7,572	0.66	Y
8	16,724,979	34,391,059	Deletion	1.1	1,607	904	0.56	9,461	6,085	0.64	Y
8	69,514,845	146,364,022	Amplification	2.8	1,623	1,964	1.21	9,677	13,603	1.41	N
9	140,363,011	140,435,225	Amplification	4.0	908	1,390	1.53	3,306	5,828	1.76	N
10	86,029,818	99,706,269	Deletion	1.3	1,591	913	0.57	9,493	6,282	0.66	Y
10	98,603,610	98,670,456	Deletion	0.6	1,849	389	0.21	11,585	2,390	0.21	Y
12	145,809	1,408,212	Amplification	2.6	1,571	1,701	1.08	8,984	11,007	1.23	N
14	36,950,751	37,431,290	Amplification	2.7	1,619	1,897	1.17	9,416	13,424	1.43	Y
14	67,021,632	67,270,506	Deletion	1.5	1,786	1,231	0.69	11,436	8,938	0.78	Y
15	20,000,078	102,432,398	Deletion	1.3	1,465	818	0.56	8,418	5,409	0.64	N
15	34,710,279	34,819,363	Deletion	0.9	688	113	0.16	3,668	720	0.20	Y
15	62,824,643	62,966,870	Deletion	0.8	1,620	395	0.24	9,642	2,478	0.26	Y
15	102,432,538	102,521,293	Amplification	3.1	520	374	0.72	1,415	1,191	0.84	N
16	5,553,164	7,351,018	Deletion	1.5	1,629	1,158	0.71	9,858	7,939	0.81	Y
16	6,541,933	6,604,871	Deletion	0.7	1,644	627	0.38	10,071	4,354	0.43	Y
17	349	22,249,120	Deletion	1.4	1,322	722	0.55	7,072	4,447	0.63	N
18	112,698	78,017,072	Deletion	1.3	1,554	874	0.56	9,102	5,878	0.65	N
20	14,752,676	15,234,961	Deletion	1.5	1,670	1,155	0.69	9,768	8,269	0.85	Y
22	16,062,595	16,475,093	Amplification	2.6	273	199	0.73	623	498	0.80	N
22	16,456,492	51,239,045	Deletion	1.4	1,212	661	0.55	6,161	3,879	0.63	N

Supplementary Table 13. Comparison of different phasing approaches.

	Barcode library described by Zheng et al.	Peters et. al.[3]	Amini et. al.[4]	de Vree et. al.[5]	Regan et. al.[6]	Borgstrom et. al. [7]
Approach	Partition genomes to droplets and construct sequencing libraries to phase and call structural variants of whole genome or exome.Libraries are compatible with Illumina sequencers.	High molecular genomic DNA is separated into 384 wells, then barcoded during library construction for sequencing on Complete Genomics platform.	Combinatorially index genomic DNA with tranposase and PCR. Libraries are compatible with Illumina sequencers.	Use proximity ligation to link genetic loci that are spatially close. Capture fragments of interest for NGS library construction and sequencing on Illumina sequencers.	Partition alleles into droplets, and use digital droplet PCR and allele-specific fluorescence probes to detect phasing of alleles.	Use emulsion compartmentalization to barcode single DNA molecules, and demonstrate phasing of bacterial 16S sequences.
Targeted?	Genome-wide, but comopatible with targetting	Genome-wide	Genome-wide	Yes	Yes, one SNP pair per well	Yes
# of partitions	>100K droplets	384 wells	10K virtual compartments	N/A	20K droplets	2M droplets*
Input	~1ng human genomic DNA (~60 molecules per bead, at 50 kb)	100pg of human DNA	100ng high molecular weight human genomic DNA (average size of 100-200kb)	100000g pellets containing crude nuclear extracts	high molecular weight DNA of 10-20ng	1 molecule per bead
Sequencing requirement	~30X	~80X	~40X	N/A	N/A	N/A
Phasing performance	phase >97% SNPs, N50=0.9-2.8Mb	phase >90% heterozygous SNVs, N50=0.5-1.6Mb	phasing of >95% of heterozygous variants, N50=1.4-2.3Mb	phase >98% SNVs of BRCA1 allele	can phase long genomic distances up to 200kb	phasing of 16S rRNAs

*M = million

Supplementary Note 1. Phasing linked-reads

The set of mapped reads for each barcode is clustered into groups such that each group has no gap between neighboring reads larger than 50 kb. These groups are very likely to originate from a single input molecule (>99.5%). Each read group is assigned a molecule index f. We record the Phred score and molecule index for each read supporting each allele of heterozygous variants. Variant phasing is determined by finding a phasing configuration of heterozygous variants that maximizes the likelihood of the observed reads and associated molecular indices (1). The following equation details the likelihood algorithm that we employed.

$$P(O|X) = \prod_f P(O_{1,f}, \dots, O_{N,f}|X) \quad (1)$$

Where

- $P\left(O_{1,f}, \dots, O_{N,f}\middle|X\right) = \frac{1-\propto}{2}(\prod_i P(O_{i,f}|A_{i,f}) + \prod_i P(O_{i,f}|A_{i,1-Xi})) + \propto \prod_i 0.5$,
- $\log P\left(O_{i,f}\middle|A_{i,p}\right) = \sum_r 1\left(S_r = A_{i,p}\right)\left(1 - 10^{-\frac{Q_r}{10}}\right) + 1\left(S_r \neq A_{i,p}\right)\left(10^{-\frac{Q_r}{10}}\right)$,
- $O_{i,f}$= observed read and barcode support at variant i, from molecule f,
- $A_{i,p}$= allele on phase p at variant i,
- X_i= phasing of variant i,
- S_r= $A_{i,p}$: read r matches allele $A_{i,p,}$
- Q_r = phred score of read r,
- α = allele collision probability, and
- Log_{10} is used.

The search for the maximum-likelihood phasing configuration is organized as follows: first, we find near-optimal local haplotype configurations with a beam search algorithm over blocks of

~40 adjacent variants. Second, the relative phasing of the blocks is determined with a greedy sweep over the block junctions. Third, we invert the haplotype assignment of individual variants to find local improvement to the phasing, and iterate until convergence. Last, the phasing configuration is broken into phase blocks that have a high probability of being internally correct. We determine the breakpoint of a phase block at each variant; this process involves comparing the log-likelihoods of the optimal configuration with a configuration where all variants to the right of the current variant have their haplotype assignment inverted (an empirically derived threshold of 0.995 was used). The confidence of the phasing assignment of single variants is computed as the log-likelihood ratio between the optimal configuration and optimal configuration with the test variant inverted.

We use a Phred-like phasing score to quantify the phasing quality at each variant, with

$$PQ_i = -10\log\left(\frac{P(data|<best\ solution\ with\ variant\ i\ flipped>)}{P(data|<best\ solution>)}\right) \quad (2)$$

Where

- $P(data|solution) = likelihood\ function\ modeling\ the\ observed\ data,$
- $< best\ solution >= final\ phasing\ result\ that\ maximizes\ the\ likelihood,$ and
- Log_{10} is used.

To evaluate the performance of our phasing algorithm, we calculated the following metrics:

- %SNVs phased = fraction of heterozygous SNVs from input VCF that were confidently phased
- % genes phased (< 100 kb) = fraction of genes with total genomic size less than 100kb

where all SNVs were contained in the same phase block

- N50 phase block = phase block size X such that half the phased genome is covered by phase blocks longer than X, and half the phased genome is covered by blocks shorter than X. A phase block is contiguous set of variants whose relative phasing has been determined by the algorithm, whereas phase block size is defined as the genomic distance between the first and last variant in the block.
- Longest phase block = the single longest phase block in the results

As additional assessment of phasing performance, we plotted the probability of any two SNVs being correctly phased over the distance between the SNVs. All phased SNVs in each phase block were considered in this analysis.

The lower the frequency of SNVs, the fewer reads and barcode support the SNVs will have. In order to phase 2 SNVs, there needs to be at least a read covering each SNV, and the reads need to have the same barcode. The lower the number of linked-reads per molecule, the less likely such rare heterozygous variants will be phased. This can be improved if (a) sequencing coverage goes up, (b) input DNA amount decreases, or (c) the input molecule size increases.

We calculated switch errors by comparing our phasing results to phasing ground-truth dataset as listed in **Table 1**. We compute short and long switch errors as by previously reported[4]. Briefly, short switch errors are individual variants with incorrect phasing and long switch errors are positions where the relative phasing of variants before the position is incorrect compared to variants after that position. Phasing errors are decomposed into short and long switch errors with a Viterbi recursion[4] that scores short switch errors as -1 and long switch errors as -5, and finds an error assignment that maximizes the total score. Switch error rates are reported as the error rate per variant with ground truth phasing available.

Ground truth comparison data originated from the following public data sets as reported by Cleary *et al.*[8] and Kitzman *et al*.[9]:

1.) ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/RTG//phasing_annotated.vcf.gz
2.) ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/RTG//ppp_full_cohort_withDNP.merged_avr0.15.vcf.gz
3.) ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/RTG//family_3.merged_avr0.15.vcf.gz
4.) http://krishna.gs.washington.edu/indianGenome/indian_snps_phased.vcf

Supplementary Note 2. Structural variant calling from linked-read data

A series of different approaches were used depending on the origin of the Linked-Reads from either whole genome versus exome approaches.

a) Whole genome sequencing data. To call large-scale structural variants, we bin the genome into 10 kb windows and count the different barcodes of mapping quality Q60 reads within each window. We use a binomial test to find all pairs of regions that are at least 50 kb apart (or on different chromosomes) and share more barcodes than what would be expected by chance (using a p-value cutoff of 10^{-15} without any multiple hypothesis correction). We found that this cutoff was loose enough to include all interesting regions of potential structural variation. The number of pairs of genomic loci that we need to compare at this step is roughly in the order of 10^{10}. In order to perform these comparisons efficiently, we encode the set of barcodes in each

genomic window as non-zero entries in a (very sparse) matrix and use sparse matrix multiplications to identify regions with overlaps.

This procedure allows us to quickly identify candidate regions for structural variation. However, the binomial test generates a very large number of false positives since it does not account for many aspects of the system, such as the length distribution of the library molecules and the variation in the amplification rate across GEMs. In a second pass, we use a probabilistic approach to clean up this initial candidate list.

First, we obtain an estimate of the set of library molecules by joining nearby reads (within 30 kb) with the same barcode. In the following discussion, we will use the term "fragment" to refer to a span of nearby reads with the same barcode. Fragments originate from some unobserved molecules (that may be longer than the observed fragments). Based on the set of fragments, we estimate quantities such read generation rate (sequenced reads per bp) of individual GEMs, the number of molecules inside each partition, and the molecule length distribution.

Given a pair of candidate windows W_1, W_2, we find the sets of fragments that overlap them and then identify pairs of fragments in W_1 and W_2 with the same barcode. Such pairs are potentially evidence for structural variation, since they suggest that the same molecule might have spanned two relatively distant loci of the genome. To quantify this evidence, we compute the following likelihood ratio score:

$$LR = \frac{P(observed\ fragments \mid SV)}{P(observed\ fragments \mid no\ SV)} \quad (3)$$

Since fragments with different barcodes are independent, this score decomposes to a product of terms with one term for each of the pairs of fragments with the same barcode b:

$$\frac{P(r_1,r_2,l_1,l_2,d \mid SV;a_b)}{P(r_1,r_2,l_1,l_2,d \mid no\ SV;a_b)} \quad (4)$$

where:

- r_1, r_2 are the number of reads on each of the two fragments,
- l_1, l_2 are the observed lengths of the two fragments,
- *d* is the distance between the two fragments, and
- a_b is the rate (reads/bp) of the GEM/barcode *b*.

The two candidate fragments might have originated from the same molecule or from different molecules, therefore:

$$P(r_1, r_2, l_1, l_2, d \mid SV; a_b) =$$
$$P(r_1, r_2, l_1, l_2, d \mid same\ molecule, SV; a_b)P(same\ molecule \mid SV) +$$
$$P(r_1, r_2, l_1, l_2, d \mid different\ molecules, SV; a_b)P(different\ molecules \mid SV) \quad (5)$$

The probability assuming that the fragments originated from different molecules is:

$$P(r_1, r_2, l_1, l_2, d \mid different\ molecules, SV; a_b) = P_{frag}(r_1, l_1; a_b)P_{frag}(r_2, l_2; a_b) \quad (6)$$

where $P_{frag}(r, l; a_b)$ is the probability of observing r reads from a molecule of unknown length such that the reads span an observed length of l.

Assuming that the reads are generated from a Poisson process with constant rate across the genome, the following is calculated:

$$P_{frag}(r, l; a_b) = \sum_{m:m\geq l}\left(r(r-1)\left(\frac{l}{m}\right)^{r-2}\frac{m-l}{m^2}\right)P_p(r; ma_b)P_L(m) = \sum_{m:m\geq l}(m-l)\,P_p(r-2; a_b l)P_p\big(0; a_b(m-l)\big)a_b^2 P_L(m) \quad (7)$$

where $P_p(r; b)$ is the probability mass function of a Poisson distribution with parameter b and $P_L(m)$ is the (pre-estimated) probability that the true molecule length is m.

The probability given that the fragments came from the same molecule can be computed in a similar way as:

$$\sum_{m:m\geq l_1+l_2+d}(m-l_1-l_2-d)\,P_p(r_1-2; a_b l_1)P_p(r_2-2; a_b l_2)P_p\big(0; a_b(m-l_1-l_2)\big)a_b^4 P_L(m) \quad (8)$$

In the presence of an SV, the likelihood is similar to equation (3). However, in this case, there is an additional unknown, namely the exact position of the breakpoints with respect to the observed fragments. For instance, assume that there was a deletion between positions 100,000 and 200,000 of chromosome 1 and that the observed fragments span the regions 85,000 - 90,000 and 210,000 - 230,000. If we knew the exact breakpoints, we could use the previous calculations with d set to 10 kb + 10 kb = 20 kb. Since the position of the true breakpoints (and therefore the true distance between the observed fragments) is unknown, we obtain an estimate of d by computing the largest extent d' such that $P_p(0; a_b d') > 0.75$. Then, we set d = 2d' and proceed as described before.

b) Targeted sequencing data. In the case of targeted sequencing such as exome linked-reads, we need to account for the composition of the target set. We assume that the off-target regions

generate reads following a similar Poisson process as the target regions, but with a different rate. In particular, let b_t be the fraction of reads on target and g_t be the fraction of the genome that is covered by the target regions. If a_b is the Poisson rate (related to barcode b) of target regions, then the rate of off-target regions is:

$$\tilde{a}_b = \frac{1-b_t}{1-g_t}\frac{g_t}{b_t}a_b \quad (9)$$

The probability of observing r reads from a region that contains l_t bp of targets and l_n bp of off-target regions is:

$$P_{mixture}(r, l_1, l_2; a_b, \tilde{a}_b) = \sum_{n=0}^{n=r} P_p(r-n; a_b l_t) P_p(n; \tilde{a}_b l_n) \quad (10)$$

The probability of observing r reads from a molecule of unknown length that spanned an observed length of $l = l_t + l_n$ is:

$$\sum_{m:m\geq l} P_{mixture}(r-2, l_t, l_n; a_b, \tilde{a}_b) a_b^2 P_L(m) \sum_{f\in offsets} P_{mixture}\left(0, d_{f_t}, d_{f_n}; a_b, \tilde{a}_b\right) \quad (11)$$

where the inner sum is taken over all $m - l$ offsets of the unobserved molecule with respect to the observed fragment, and d_{f_t} and d_{f_n} are the bases on and off-target for the corresponding offset. To simplify calculations, for a given value of m, we compute the average fraction of bases on and off-target across all offsets and assume that all offsets have the same target composition.

The rest of the probabilities needed to compute (2) are adjusted in a similar way from the WGS case. In practice all probabilities were computed in log-space to avoid underflows. We used a

log-likelihood ratio cutoff of 200. We empirically found that this cutoff resulted in high-quality calls with very low false positive rates after the filtering steps described below.

c) Refining breakpoints using short read information. After obtaining breakpoint windows using the approach described above, we used information from read pairs and split reads to further refine the breakpoint locations. For each called structural variant, we selected all read pairs and split reads within the called breakpoint windows. We used a probabilistic approach similar to a previous study[10] to infer the breakpoint loci based on the combined evidence from all selected read pairs and split reads. In order to avoid false positives, we only attempted to infer the exact breakpoint loci when there were at least 4 read pairs and split reads supporting the call.

d) Filtering calls based on gaps and segmental duplications. We excluded SV calls whose breakpoints overlap different copies of the same segmental duplication (using the Segmental Duplication track from the UCSC browser). Structural variation is enriched in such regions[11], so some of these calls might represent true events. However, we noticed that a large fraction of calls in regions of structural variation are the result of the inability of aligners to properly resolve repetitive regions, since a small amount of variation is sufficient to make reads map uniquely and with high mapping quality to one or the other copy of the segmental duplication.

We further excluded SV calls that are within 10 kb from gaps (using the gaps track from the UCSC browser) or from new sequence introduced in hg38 (using the hg19 diff track from the UCSC browser). SV calls in these regions are likely related to assembly errors in constructing the hg19 reference builds.

References

1. Pendleton, M. et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* (2015).
2. Kidd, J.M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
3. Peters, B.A. et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190-195 (2012).
4. Amini, S. et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nature genetics* **46**, 1343-1349 (2014).
5. de Vree, P.J. et al. Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nature biotechnology* **32**, 1019-1025 (2014).
6. Regan, J.F. et al. A rapid molecular approach for chromosomal phasing. *PloS one* **10**, e0118270 (2015).
7. Borgstrom, E. et al. Phasing of single DNA molecules by massively parallel barcoding. *Nature communications* **6**, 7173 (2015).
8. Cleary, J.G. et al. Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. *Journal of computational biology : a journal of computational molecular cell biology* **21**, 405-419 (2014).
9. Kitzman, J.O. et al. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature biotechnology* **29**, 59-63 (2011).
10. Layer, R.M., Chiang, C., Quinlan, A.R. & Hall, I.M. LUMPY: a probabilistic framework for structural variant discovery. *Genome biology* **15**, R84 (2014).
11. Mills, R.E. et al. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65 (2011).