Supplementary Information for: "Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event"

Tanya Golubchik[1]*, Angela B. Brueggemann[2]*, Teresa Street[1], Robert E. Gertz Jr.[3], Chris C. A. Spencer[4], Thien Ho[1], Eleni Giannoulatou[4], Ruth Link-Gelles[3], Rosalind M. Harding[2], Bernard Beall[3], Tim E. A. Peto[5], Matthew R. Moore[3], Peter Donnelly[4]†, Derrick W. Crook[5]†, Rory Bowden[1,4,5]†

* These authors made equal contributions to the work. † These authors jointly supervised the work.

Correspondence to: Peter Donnelly (peter.donnelly@well.ox.ac.uk)

[1] Department of Statistics, University of Oxford, Oxford, UK;

[2] Department of Zoology, University of Oxford, Oxford, UK;

[3] National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA;

[4] Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK;

[5] Oxford Biomedical Research Centre, Nuffield Department of Medicine, University of Oxford, Oxford, UK.

Supplementary Table 1: Sample Details

Group	Sample ID	State	Year	Serotype	MLST allelic profile *aroE*	*gdh*	*gki*	*recP*	*spi*	*xpt*	*ddl*	ST	CC
Recombinant - P1 (n = 42)													
	cdc10	CT	2003	19A	16	13	4	4	6	113	18	695	247
	cdc12	NY	2003	19A	16	13	4	4	6	113	18	695	247
	cdc13	NY	2003	19A	16	13	4	4	6	113	18	695	247
	cdc780	NY	2003	19A	16	13	4	4	6	113	18	695	247
	cdc14	CT	2004	19A	16	13	4	4	6	113	18	695	247
	cdc16	NY	2004	19A	16	13	4	4	6	113	18	695	247
	cdc782	NY	2004	19A	16	13	4	4	6	113	18	695	247
	cdc24	CT	2005	19A	16	13	4	4	6	113	18	695	247
	cdc27	MN	2005	19A	16	13	4	4	6	113	18	695	247
	cdc33	CT	2005	19A	16	13	4	4	6	113	18	695	247
	cdc34	CT	2005	19A	16	13	4	4	6	113	18	695	247
	cdc37	CT	2005	19A	16	13	4	4	136	113	18	2363	247
	cdc39	MD	2005	19A	16	13	4	4	6	113	18	695	247
	cdc42	NY	2005	19A	16	13	4	4	6	113	18	695	247
	cdc44	NY	2005	19A	16	13	4	4	6	113	18	695	247
	cdc49	CT	2005	19A	16	13	4	4	6	113	18	695	247
	cdc108	MN	2005	19A	16	13	4	4	6	113	18	695	247
	cdc707	GA	2005	19A	16	13	4	4	6	113	18	695	247
	cdc75	GA	2006	19A	16	13	4	4	6	113	18	695	247
	cdc78	MD	2006	19A	16	13	4	4	6	113	18	695	247
	cdc79	MD	2006	19A	16	13	4	4	6	113	18	695	247
	cdc81	MD	2006	19A	16	13	4	4	6	113	18	695	247
	cdc90	NY	2006	19A	16	13	4	4	6	113	18	695	247
	cdc91	CO	2006	19A	16	13	4	4	6	113	18	695	247
	cdc128	CT	2007	19A	16	13	4	4	6	113	18	695	247
	cdc138	CT	2007	19A	16	13	4	4	6	113	18	695	247
	cdc195	NY	2007	19A	16	13	4	4	6	113	18	695	247
	cdc212	CT	2007	19A	16	13	4	4	6	113	18	695	247
	cdc248	NY	2007	19A	16	13	4	4	6	113	18	695	247
	cdc386	GA	2007	19A	16	13	4	4	6	113	18	695	247
	cdc406	MN	2007	19A	16	13	4	4	6	113	18	695	247
	cdc428	NY	2007	19A	16	13	4	4	6	113	18	695	247
	cdc460	CO	2007	19A	16	13	4	4	6	113	18	695	247
	cdc481	GA	2007	19A	16	13	4	4	6	113	18	695	247
	cdc543	CT	2007	19A	16	13	4	4	6	113	18	695	247
	cdc577	NY	2007	19A	16	13	4	4	6	113	18	695	247
	cdc596	CT	2007	19A	16	13	4	4	6	113	18	695	247
	cdc601	TN	2007	19A	16	13	4	4	6	113	18	695	247
	cdc622	OR	2007	19A	16	13	4	4	6	113	18	695	247
	cdc653	MD	2007	19A	16	13	4	4	6	113	18	695	247
	cdc687	MN	2007	19A	16	13	4	4	6	113	18	695	247
	cdc741	MD	2007	19A	16	13	4	4	6	113	18	695	247

Supplementary Table 1 (cont.): Sample Details

Group	Sample ID	State	Year	Serotype	MLST allelic profile *aroE*	*gdh*	*gki*	*recP*	*spi*	*xpt*	*ddl*	ST	CC
Recombinant - P2 (n = 8)													
	cdc25	NY	2005	19A	16	13	4	5	6	197	14	2365	247
	cdc26	NY	2005	19A	16	13	4	5	6	197	14	2365	247
	cdc35	CT	2005	19A	16	13	4	5	6	197	14	2365	247
	cdc41	NY	2005	19A	16	13	4	5	6	197	14	2365	247
	cdc250	CT	2007	19A	16	13	4	5	6	197	14	2365	247
	cdc251	CT	2007	19A	16	13	4	5	6	197	14	2365	247
	cdc305	GA	2007	19A	16	13	4	5	6	197	14	2365	247
	cdc360	NY	2007	19A	16	13	4	5	6	197	14	2365	247
Other Recombinants (P3, P4, P5) (n = 3)													
	cdc21	CO	2005	19A	16	13	4	4	6	10	18	899	247
	cdc54	CO	2006	19A	16	13	4	4	6	113	18	695	247
	cdc570	NY	2007	19A	16	13	4	4	6	113	18	695	247
Prospective Donors (n = 5)													
	cdc02	NY	1999	19A	8	13	14	4	17	4	14	199	199
	cdc08	GA	2003	19A	8	13	14	4	17	4	14	199	199
	cdc09	GA	2003	19A	8	13	14	4	17	4	14	199	199
	cdc11	GA	2003	19A	8	13	14	4	17	4	14	199	199
	cdc06	CT	1999	19A	7	13	14	4	17	4	14	645	199
Prospective Recipients (n = 4)													
	cdc01	CT	1999	4	16	13	4	4	6	113	18	695	247
	cdc03	NY	1999	4	16	13	4	4	6	113	18	695	247
	cdc04	CT	1999	4	16	13	4	4	6	113	18	695	247
	cdc07	MD	2002	4	16	13	4	4	6	10	18	899	247

Supplementary Table 2: Capsular Imports

Recombinant lineage	Isolates sequenced	Isolates available	donor-like sequence around *cps*	Size of observed imports, 5'-*cps* (kb)	Size of observed imports, 3'-*cps* (kb)	Inferred total size of cps import(s) (kb)*
P1	42	175	306862 – 348602	15.4	9.4	39.4
P2	8	8	300206 – 352813	22.1	13.6	50.3
P3	1	1	296948 – 302696, 309361 – 322278	5.7, 12.9	–	27.5 – 39.9
P4	1	1	294281 – 341296, 342627 – 345492, 349562 – 353096	28.2	2.1, 2.9, 3.5	44.7 – 56.5
P5	1	1	312511 – 341296	10.0	2.179	26.5

* size including flanking sequences, assuming transfer of complete canonical 19A *cps*

Supplementary Table 3: Additional Heterologous Fragments

Fragments of the genome identified as imports in each vaccine escape recombinant lineage

Region	Start	End	SNPs	No. array fragments spanned	Size (kb)
(P1)					
p1r0	303865	305968	5	1	2.104
19A *cps*	306862	348602	168	7	41.741
p1r1	1699614	1700503	6	1	0.89
p1r2	1767941	1769021	4	1	1.081
p1r3	1780912	1788180	13	2	7.269
(P2)					
p2r1	66093	69956	3	3	3.864
p2r2	287564	289916	15	1	2.353
p2r3	289921	290201	4	1	0.281
19A *cps*	300206	346820	168	8	46.615
p2r4	772306	816939	12	5	44.634
p2r5, total extent	1479144	1574205	26	4	95.062
Part 1	1479144	1480380	3	1	1.237
Part 2	1547899	1574205	23	3	26.307
p2r6	1630427	1630949	5	1	0.523
p2r7	1749323	1777793	32	7	28.471
p2r8	2104348	2105004	6	1	0.657
(P3)					
p3r1	69034	69307	3	1	0.274
p3r2	261799	262255	3	1	0.457
p3r3	289943	294357	5	2	4.415
p3r4	295493	296546	3	1	1.054
19A *cps*	296948	302696	21	2	5.749
19A *cps*	309361	322278	57	5	12.918
p3r5	339468	339953	3	1	0.486
p3r6	341116	341997	5	1	0.882
p3r7	342165	342763	5	1	0.599
p3r8	343299	343986	2	1	0.688
p3r9	345172	345984	7	1	0.813
p3r10	346062	346509	3	1	0.448
p3r11	349820	350324	5	1	0.505
p3r12	801026	801746	2	1	0.721
p3r13	802058	802522	2	1	0.465
p3r14	923423	943833	12	3	20.411
p3r15	951784	952459	3	1	0.676
p3r16	1352857	1355029	3	2	2.173
p3r17	1355525	1355621	2	1	0.097
p3r18	1479798	1480347	3	1	0.55
p3r19	1630427	1632094	16	1	1.668
p3r20	1652648	1654228	8	1	1.581
p3r21	1743603	1745044	4	1	1.442
p3r22	1840713	1866523	2	2	25.811
p3r23	1895863	1916353	4	2	20.491
p3r24	1916640	1916797	2	1	0.158
p3r25	1917066	2009033	23	3	91.968
p3r26	2103857	2104412	3	1	0.556
p3r27	2104592	2105004	4	1	0.413
(P4)					
p4r1	165798	165834	2	1	0.037
19A *cps*	294281	341296	156	7	47.016
19A *cps*	342627	345492	19	1	2.866
19A *cps*	349562	353096	17	1	3.535
(P5)					
p5r1	211393	214427	20	1	3.035
p5r2	287564	290323	21	1	2.76
19A *cps*	312511	341296	57	5	28.786
p5r3	1314865	1318996	18	2	4.132

Supplementary Table 4: Spread of P1

Incidence of serotype 19A, P1 recombinants (CC695), CC320, and other serotype 19A strains, by ABCs site, under 5 year-olds, 2003-2007. Absolute counts (% of state total). Percentages are the data in Figure 2 in the main paper. "-": data were not collected in New Mexico before 2004.

	State	CA	CO	CT	GA	MD	MN	NM	NY	OR	TN	Total
2003	tested n	37 (100)	37 (100)	42 (100)	99 (100)	33 (100)	71 (100)	- (-)	24 (100)	16 (100)	48 (100)	407 (100)
	all 19A	7 (19)	5 (14)	7 (17)	24 (24)	7 (21)	12 (17)	- (-)	6 (25)	5 (31)	4 (8)	77 (19)
	P1	0 (0)	0 (0)	1 (2)	0 (0)	0 (0)	0 (0)	- (-)	2 (8)	0 (0)	0 (0)	3 (1)
	CC320	0 (0)	0 (0)	3 (7)	0 (0)	1 (3)	1 (1)	- (-)	0 (0)	0 (0)	0 (0)	5 (1)
	other 19A	7 (19)	5 (14)	3 (7)	24 (24)	6 (18)	11 (15)	- (-)	4 (17)	5 (31)	4 (8)	69 (17)
2004	tested n	27 (100)	24 (100)	30 (100)	68 (100)	32 (100)	74 (100)	18 (100)	23 (100)	15 (100)	62 (100)	373 (100)
	all 19A	5 (19)	5 (21)	9 (30)	26 (38)	11 (34)	18 (24)	9 (50)	7 (30)	4 (27)	22 (35)	116 (31)
	P1	0 (0)	0 (0)	1 (3)	0 (0)	0 (0)	0 (0)	0 (0)	2 (9)	0 (0)	0 (0)	3 (1)
	CC320	0 (0)	0 (0)	1 (3)	3 (4)	4 (13)	2 (3)	0 (0)	0 (0)	0 (0)	0 (0)	10 (3)
	other 19A	5 (19)	5 (21)	7 (23)	23 (34)	7 (22)	16 (22)	9 (50)	5 (22)	4 (27)	22 (35)	103 (28)
2005	tested n	19 (100)	28 (100)	43 (100)	77 (100)	44 (100)	79 (100)	22 (100)	21 (100)	13 (100)	51 (100)	397 (100)
	all 19A	6 (32)	11 (39)	19 (44)	28 (36)	17 (39)	23 (29)	5 (23)	12 (57)	2 (15)	13 (25)	136 (34)
	P1	0 (0)	0 (0)	1 (2)	0 (0)	1 (2)	1 (1)	0 (0)	3 (14)	0 (0)	0 (0)	6 (2)
	CC320	0 (0)	1 (4)	6 (14)	6 (8)	8 (18)	9 (11)	0 (0)	1 (5)	0 (0)	5 (10)	36 (9)
	other 19A	6 (32)	10 (36)	12 (28)	22 (29)	8 (18)	13 (16)	5 (23)	8 (38)	2 (15)	8 (16)	94 (24)
2006	tested n	23 (100)	27 (100)	45 (100)	63 (100)	62 (100)	69 (100)	27 (100)	27 (100)	8 (100)	48 (100)	399 (100)
	all 19A	5 (22)	11 (41)	22 (49)	34 (54)	27 (44)	21 (30)	11 (41)	16 (59)	5 (63)	21 (44)	173 (43)
	P1	0 (0)	0 (0)	6 (13)	3 (5)	1 (2)	0 (0)	0 (0)	1 (4)	0 (0)	0 (0)	11 (3)
	CC320	1 (4)	3 (11)	10 (22)	5 (8)	9 (15)	5 (7)	4 (15)	4 (15)	1 (13)	4 (8)	46 (12)
	other 19A	4 (17)	8 (30)	6 (13)	26 (41)	17 (27)	16 (23)	7 (26)	11 (41)	4 (50)	17 (35)	116 (29)
2007	tested n	30 (100)	16 (100)	38 (100)	79 (100)	39 (100)	96 (100)	32 (100)	31 (100)	15 (100)	61 (100)	437 (100)
	all 19A	13 (43)	4 (25)	21 (55)	44 (56)	21 (54)	36 (38)	8 (25)	11 (35)	4 (27)	25 (41)	187 (43)
	P1	1 (3)	1 (6)	9 (24)	6 (8)	3 (8)	1 (1)	0 (0)	3 (10)	0 (0)	1 (2)	25 (6)
	CC320	4 (13)	0 (0)	9 (24)	13 (16)	8 (21)	15 (16)	2 (6)	3 (10)	0 (0)	11 (18)	65 (15)
	other 19A	8 (27)	3 (19)	3 (8)	25 (32)	10 (26)	20 (21)	6 (19)	5 (16)	4 (27)	13 (21)	97 (22)

Supplementary Table 5: Primers for amplifying capsular recombination breakpoints

Region	Primer name	Primer sequence
cps upstream	*cps* _up_F	GCGGATGAATCAGGATGC
	cps _up_R	CGCAGAGTTGGAAAAAGTCT
cps downstream	*cps* _down_F	CGTGAACGATAGTAGCAGTTGA
	cps _down_R	GCCTTAACCAAATCTGTGGGAATATC

Supplementary Table 6: Sequencing performance

Metric	Raw (Affymetrix)	Filtered
Call rate	96.4%	91.8%
No-Call rate	3.6%	8.2%
Error rate	0.1%	0.004%
Sensitivity (differences from reference detected)	72.3%	49.2%
Specificity (differences detected correctly)	87.9%	99.4%

Supplementary Table 7: Primers for PCR-RFLP assays

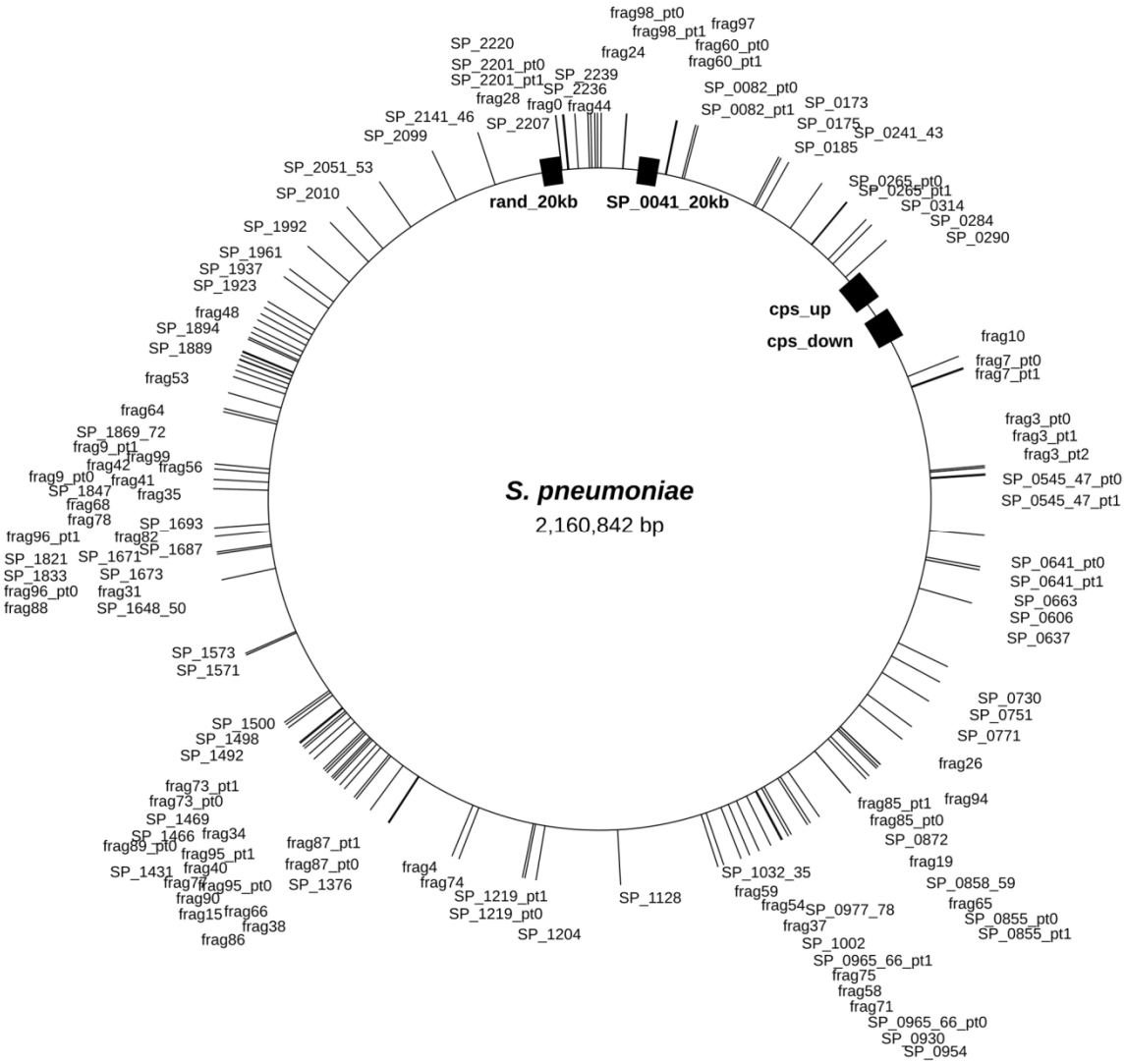
Region	Primer name	Primer sequence	Position(s) typed	Restriction enzyme
(P1)				
p1r0	p1_r0-F	GTCCTTTATTGAGCGCAGGAAG	304390	BfuCI
	p1_r0-R	CTGGTCTGAAAGCAGGTTCGAT		
p1r1	p1_r1-F	GTAAAGCGGTGCTGTTTCTGCT	1699748	BamHI
	p1_r1-R	AGTCCAAGTGACCATCGCATCT		
p1r2	p1_r2-F	TGCAAGCCCTTATTTCCATCTT	1767941, 1768401, 1768407	BstBI, BsrDI, NheI
	p1_r2-R	TTGATTATCGGTACGGCAACCT		
p1r3	p1_r3_1-F	GTTCGATGACTTCCGTAATGGG	1780936, 1780972	AciI, NlaIII
	p1_r3_1-R	CATGGTAGGTGAGCGGATTTCT		
	p1_r3_2-F	AGCAATTCAACTCCCTCGTCAG	1787249	AciI
	p1_r3_2-R	CTATTGGTATCGGTGGTTTGCC		
(P2)				
p2r1	p2_r1snp2-F	CGCTTATGACTTTACAGTCGAAGAG	66649	AciI
	p2_r1snp2-R	CGTCGTCTACAACCATAATGCG		
	p2_r1snp6-F	GATCTTCTCGGTGGGACAC	69307	BsrDI
	p2_r1snp6-mis-R	AACAGGTCATTAACTTTCGCAAT		
p2r2	p2_r2snp8_9-F	CTCTTAAGGAAGGGACAACCA	288334	BfaI
	p2_r2snp8_9-R	CAAGGTATTGTTGATGGCACGA		
	p2_r2snp13-F	CTCTAGGTGGAAACTTAGTTG	289196	AluI
	p2_r2snp13-R	ATATCCGCTATTCGGTGAATCC		
p2r3	p2_r3snp3-F	AGACGGATGCTAAGAGAGCG	289921, 290092	TaqI, NlaIII
	p2_r3snp3-R	ATTGATATCCTTCCAGGCTC		
p2r4	p2_r4snp4-mis-F	ATTGCTATGCTGGCGGCTAATAT	802058	SspI
	p2_r4snp4-R	TCTCAATTGCTTTCGCAGTGTC		
	p2_r4snp16-F	GGAAGCTTGTCGAAAGACTGGA	816170	NlaIII
	p2_r4snp16-R	TGCCGTAAAGGACACCGATAC		

Supplementary Table 8: Distribution of serotype 19A-like single nucleotide variants in circulating Serotype 4 and 19A isolates

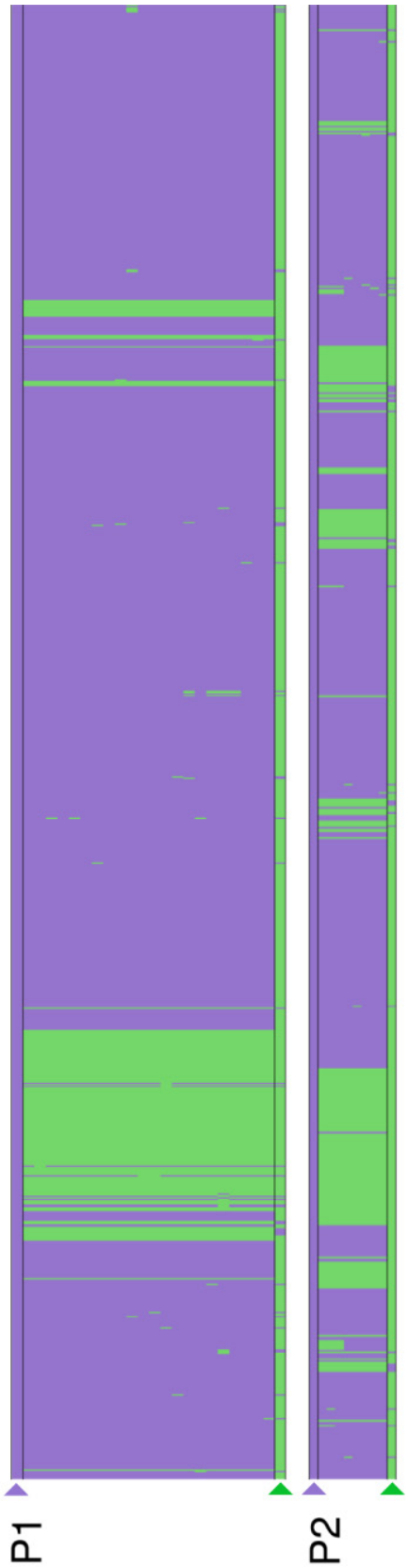
Region	Start	End	SNPs	No. array fragments spanned	Size (kb)	Serotype 4 prevalence (no. typed)	Serotype 19A prevalence (no. typed)
(P1)							
p1r0	303865	305968	5	1	2.104	0.60 (87)	0.97 (319)
p1r1	1699614	1700503	6	1	0.89	0.00 (85)	0.75 (295)
p1r2	1767941	1769021	4	1	1.081	0.00 (85)	0.75 (167)
p1r3	1780912	1788180	13	2	7.269	0.27 (71)	0.93 (275)
(P2)							
p2r1	66093	69956	3	3	3.864	0.31 (77)	0.23 (231)
p2r2	287564	289916	15	1	2.353	0.06 (77)	0.97 (177)
p2r3	289921	290201	4	1	0.281	0.13 (83)	0.65 (326)
p2r4	772306	816939	12	5	44.634	0.04 (79)	0.74 (304)

Supplementary Figure 1: Array Coverage

The Affymetrix GeneChip CustomSeq micro-array encodes 300kb of *S. pneumoniae* sequences (~12% of the genome).

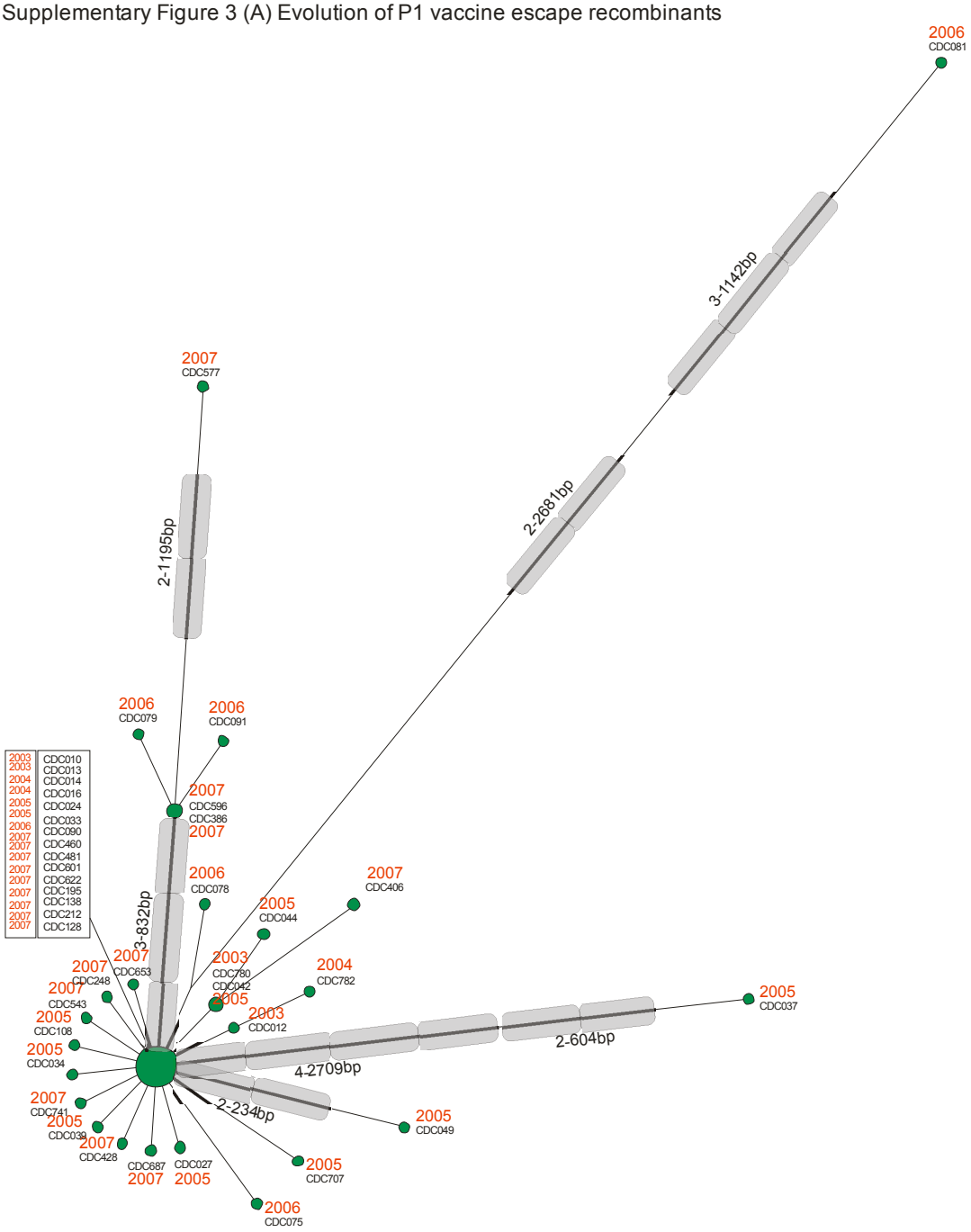


Supplementary Figure 2: Multiple-fragment recombination



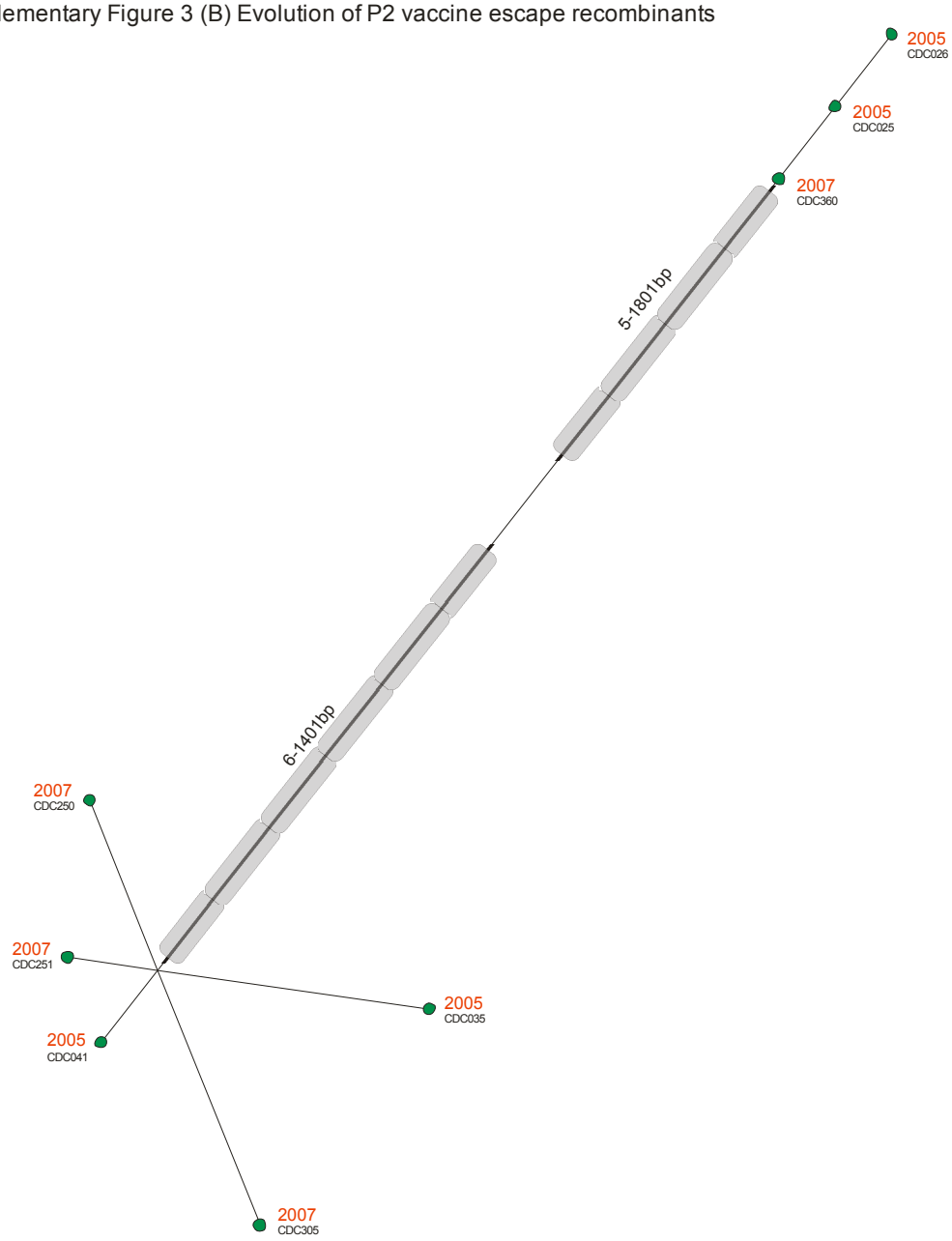
Supplementary Figure 2 Legend: Schematic in matrix form (landscape orientation) showing imports into the serotype 4 genome, for P1 (upper, 42 isolates in this figure) and P2 (lower panel, 8 isolates) lineage recombinants. Top and bottom rows of each panel (arrowed) are respectively the best-matching serotype 4 and serotype 19A isolates, with individual recombinant isolates as intervening rows. Columns are type-able sites on the chip at which differences between serotypes 4 and 19A were detected; these are coloured purple for identity to the serotype 4 parent, and green for non-serotype 4-like variants. The large mass of green colour towards the left of each panel represents the capsular (19A-like) region, and further vertical green stripes represent AHFs. Some mutations and imports have occurred since the origin of P1 or P2 (vertical green stripes that do not extend across all members of a lineage). Because different sets of sites are filtered out in each lineage and because this figure is a composite of variable sites, the horizontal scale differs between P1 and P2, and in comparisons with Main Figure 1.

Supplementary Figure 3 (A) Evolution of P1 vaccine escape recombinants



Supplementary Figure 3 Legend: Variation within a single recombinant lineage is represented as a network [1], providing an intuitive visual representation of mutation and recombination events after each initial capsular switch event. Recombination events can lead to long branches. Individual high-probability recombination tracts are represented as gray lozenges labelled with number of new variable positions introduced by the imported sequences and the length of each tract. Lengths of branches are proportional to the number of nucleotide differences, with the shortest branch equal to a single nucleotide change. Areas of nodes are proportional to the number of individuals sharing a genotype, with the smallest area equal to one isolate. (A) P1; (B) P2.

Supplementary Figure 3 (B) Evolution of P2 vaccine escape recombinants



Supplementary Note

Identification and spread of vaccine-escape recombinants

We focused on identifying a particular subset of vaccine-escape recombinants for study: those in which serotype 4 (vaccine) capsular sequences were replaced with the genes encoding (non-vaccine) serotype 19A. Invasive disease-causing serotype 19A isolates were collected by the CDC in the years 2002-2007 (n=2905) through the ABCs surveillance network, which comprises medical centres in 10 states and during the study period reached approximately 29-30 million people or 9-10% of the US population [2,3]. Serotype 19A isolates collected from 2001-2004 were genotyped by pulsed-field gel electrophoresis and a subset each year were also genotyped by MLST (n = no. of 19A genotyped/total no. of 19A available for genotyping): 2001 (n = 39/131); 2002 (n = 85/192); 2003 (n = 71/299); and 2004 (n = 89/408). In 2005 nearly all (n = 550/566) serotype 19A isolates were fully genotyped by MLST. Isolates from 2006-2007 were screened at two MLST loci (*xpt* and *ddl*): 2006 (n = 621/651); and 2007 (n = 633/658).

Putative vaccine escape recombinants from 2003-2005 have previously been reported [3,4]. (Note that one additional putative P1 progeny isolate from 2003 was identified after our initial report and is included in the current study.) To confirm the P1 progeny designation and identify any new progeny strains, 54 putative P1 progeny strains from 2006 and 658 serotype 19A isolates from 2007 were also screened by PCR amplification and sequencing of ~850 bp regions around the capsular (*cps*) locus recombinational breakpoints identified in the earliest P1 progeny (i.e. recovered from 2003-2005). Primers were designed at the P1 progeny *cps* locus breakpoints and included each breakpoint plus regions upstream and downstream, respectively, of the breakpoints. The primer sequences were as follows: *cps*_up_F (5' GCG GAT GAA TCA GGA TGC); *cps*_up_R (5' CGC AGA GTT GGA AAA AGT CT); *cps*_dn_F (5' CGT GAA CGA TAG TAG CAG TTG A); *cps*_dn_R (5' GCC TTA ACC AAA TCT GTG GGA ATA TC). Previously sequenced P1, P2 and P3 progeny strains, serotype 19A donor and serotype 4 recipient strains were used as controls.

This approach identified 186 putative recombinant isolates recovered during 2003-2007, representing five distinct capsular recombination events: P1, P2, P3 [4], P4 and P5. This same approach also determined that, for all progeny, the most likely donors had MLST-defined sequence types belonging to clonal complex (CC) 199; the most likely recipient genomes were members of CC247. Of the five progeny types, three were observed just once, and two (P1, 175 isolates, and P2, 8 isolates) were observed multiple times in ABCs.

Genomic analysis of recombinant progeny

To examine the recombinants at a level beyond the limited resolution offered by conventional sequencing, we designed a GeneChip CustomSeq array (Affymetrix, Santa Clara, CA, USA) [5,6] to re-sequence 300,000 positions (~12%) of the genome in representatives of all the recombinant lineages, plus selected serotype 4 and serotype 19A isolates representing inferred parental lineages (Supplementary Table 1). In this method, labelled, fragmented DNA from a sample is hybridised to a 'chip' carrying a series of overlapping probes comprising a known reference sequence, and the sequence is read off from the pattern of

fluorescence at successive probes. Unlike other next-generation sequencing methods, chip-based sequencing does not require assembly, as all probe positions correspond to known positions in the reference, but a high density of variants may reduce the number of positions that can be typed using a given reference sequence. We designed the re-sequencing array and filtered the data to take account of this issue, increasing accuracy and specificity for detecting departures from the reference, at the expense of per-site sensitivity.

Selection and optimisation of reference sequences for CustomSeq array

The reference sequence for fragments to be tiled on the CustomSeq array consisted of 300 kb of sequence, allocated into three categories: (i) 30 kb each up- and downstream of the *cps* locus, including 4 *cps* locus genes (*wzg*, *wzh*, *wzd*, *wze*) that are conserved across all serotypes, (ii) 100 kb of arbitrary 2 kb fragments, plus an arbitrary contiguous 20 kb sequence, and (iii) 120 kb of specific genes of interest (Supplementary Figure 1).

Arbitrary fragments were selected to give maximum coverage around the genome: the genome sequence was split into ten equal segments, and non-overlapping fragments were placed randomly on five alternate segments. Fragments from category (iii) were found on all 10 segments. Reference sequences were selected from a whole-genome alignment of TIGR4 (serotype 4, NCBI accession NC_003028) [7] against all *S. pneumoniae* genomic sequences available at the time (n = 24) (GenBank), constructed using Mauve [8]. To generate the final reference sequence, a consensus was taken from this alignment and then refined as follows: Regions of very high diversity, defined as containing 3 or more polymorphic sites in a 25 base window of the alignment, and regions absent from one or more genomes, were excluded. The remainder was subjected to an optimisation process to maximise the number of detectible polymorphic sites in the alignment. Sites were considered 'detectible' if they were flanked by at least 12 non-polymorphic sites on either side. During optimisation, consensus residues at random positions were iteratively replaced by other residues from the same position in the alignment, maintaining changes that increased the number of detectible residues. Fragments were included only if the expected base-calling rate (proportion of detectible sites) was above 80%.

CustomSeq array protocol

A single colony of each pneumococcal strain was grown overnight in 15ml Todd-Hewitt broth and DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen UK). For sequence analysis 20µg of DNA was first sheared in a sonicating water bath then reduced in size by DNase I digestion to approximately 20-200bp in length. 16µg of the fragmented DNA was labelled and then hybridised to the array according to the GeneChip CustomSeq Resequencing Array Protocol v2.1 (Affymetrix). After hybridisation the arrays were washed, stained and scanned, and the images were analysed and converted to sequence data using GeneChip Operating Software (GCOS v. 1.4, Affymetrix) and GeneChip Sequence Analysis Software (GSEQ v.4.0, Affymetrix).

Filtering CustomSeq data

Since each position on a resequencing array corresponds to a known position in the reference genome, it is trivial to construct a multiple alignment of the resequenced genomes. However, it is necessary to exclude sites for which the resequencing method can be expected to perform poorly, particularly where two or more predicted differences from the reference occur in close proximity, because hybridization at the central position of a probe may be affected by a mismatch elsewhere in the probe. We filtered the Affymetrix data by considering all such sites to be uncalled (ie substituting an 'N' at these positions); likewise, called bases within a run of Ns assigned by the Affymetrix software were replaced by 'N'. To assess the performance of such a filter we compared a subset of sequences from the array data (a total of 602620 sites) against the same sequences determined by traditional Sanger technology. Filtering reduced the error rate from 0.11% to ~ 0.004% at the expense of a reduction in the per-site call rate from 96.4% to 91.8% (Supplementary Table 6). This is equivalent to ~11 errors expected per isolate across all positions left after filtering (mean 275,740 per chip), a rate not expected to significantly affect our results.

Independent capsular recombination events define 5 lineages

All available P2, P3, P4 and P5 isolates, plus 42 of the 175 available P1 isolates, encompassing a range of locations in the US and all years in which P1s were detected, were analysed. We also resequenced isolates of serotype 4 and serotype 19A representing the inferred donor and recipient lineages (Supplementary Table 1). After filtering to remove low-confidence calls, the alignment of sequences from all recombinants and putative parents comprised a total of 767 well-supported, fully-called variable sites (SNPs). Comparison of recombinants with their prospective parents allowed us to identify capsular recombination breakpoints, confirming previous conventional sequence data, demonstrating that each progeny lineage was an entity distinct from other progeny and establishing the size range of of the *cps* imports for each lineage of recombinants at 26.5 – 56.5 kbp (Supplementary Table 2). P1 and P2 have also acquired 19A-like sequences at the two penicillin-binding protein (*pbp*) genes flanking the capsular locus – *pbp2x* and *pbp1a*, so they may have benefitted from positive selection for drug resistance as well as vaccine escape. P4 also acquired *pbp2x* along with the *cps* locus, but not *pbp1a*.

Additional Heterologous Fragments

As well as the expected *cps* imports, comparisons of variant positions identified additional fragments of DNA sequence elsewhere in the genome that matched the putative 19A donors, rather than the serotype 4 recipients (Supplementary Figure 2). We called such fragments AHFs (**a**dditional **h**eterologous **f**ragments), defined as ≥2 consecutive variants segregating between serotype 4 and 19A where all members of a recombinant lineage carried the 19A-characteristic allele. AHFs were identified that were specific to both P1 and P2 lineages, where well-matched 19A donors had been identified previously. The single detected isolate of each of P4 and P5 similarly possessed unique AHFs outside the *cps* flanks. The single isolate representing P3 had a large number of genomic fragments that did not match closely any of the resequenced serotype 4 or 19A isolates, hampering identification of prospective AHFs: either the P3 recipient was a rare, divergent serotype 4 lineage that we did not sample

in our search for parents, or P3 might be a mosaic resulting from one or more large-scale recombination episodes. The following description focuses on P1 and P2, where multiple isolates were available.

The working definition of AHF excludes 4 single, isolated variant sites in each of P1 and P2, which were considered to be more likely the result of individual SNPs segregating in the pneumococcal population. Sizes of AHFs were estimated from the distance on the reference sequence between the first and last 19A-like variant, ranging from 0.9 – 7.3 kb in P1 and 0.3 – 95.1 kb in P2 (Supplementary Table 3). Our analysis counts the minimum number of recombination fragments that could be responsible for the observed data: where an AHF appears to span several genome fragments on the re-sequencing array the true size could be smaller and the true number could be greater.

Potential origin of serotype 19A-like regions in P1 and P2

The presence of AHFs has two plausible explanations. The first is that in each lineage the capsular switch occurred on a serotype 4 background that already contained these regions. If so, none of the serotype 4 isolates we resequenced was a true recipient, since none carried these regions. Alternatively, the AHFs could have been imported along with the serotype 19A *cps* DNA during the encounter between the recipient and the donor, as part of a single series of distinct recombination events. If true, a single encounter between two pneumococcal cells in the host could be sufficient to yield mosaic progeny. Further possibilities seem less likely: for example, we cannot formally exclude that the AHFs originate in a series of co-infections. More likely is that a single co-infection event led to release of multiple genomic DNA fragments and their subsequent uptake by a transiently competent serotype 4 cell. Apart from the transfer of the *cps* locus on which progeny were ascertained, the fragments appear to have random locations (Supplementary Table 3).

We investigated the possibility that AHFs pre-existed among circulating serotype 4 genomes by using PCR-RFLP (PCR-restriction fragment length polymorphism) analysis to extensively survey serotype 4 and serotype 19A populations. We designed individual PCR-RFLP assays specific for 1-3 SNPs in each of four AHFs in each of P1 and P2 (Supplementary Table 7), and screened a pool of serotype 4 (n=88) and serotype 19A (n=344) isolates previously genotyped by the CDC around the time when the P1 and P2 recombinants emerged (1999-2005).

None of the serotype 4 isolates we studied matched either P1 or P2 across all the screened AHFs, indicating that if a plausible serotype 4 recipient carrying the AHFs existed, it must have been rare in the US population. Furthermore, each region individually was far more prevalent among the serotype 19As than among the serotype 4s, implying that a plausible parental lineage must be substantially different across the genome from known serotype 4s (Supplementary Table 8). In contrast, and as expected, several serotype 19As were found that carried the appropriate AHFs and therefore were plausible donors. Together these findings are consistent with the hypothesised multiple-transfer model.

Whole-genome sequencing of representative isolates

We used Illumina GAII next-generation sequencing technology (Illumina, San Diego, CA, USA) to confirm the pattern of recombination events in P1. Sets of 51b paired-end reads from three isolates (cdc001, cdc002, cdc013) were each assembled de novo using Velvet into contigs which were co-aligned to the TIGR4 reference genome [7] using Mauve [8]. A simple heuristic was used to identify imported sequences: intervals containing 3 or more donor-like variants interspersed with no more than 1 recipient-like variant were coloured yellow (donor) (Figure 2). Whole-genome data revealed an additional 8 recombination events that were not sequenceable by the CustomSeq array. The pattern of recombination events was also inferred using a Hidden Markov Model of the origin of sequences (from either the donor or the recipient genome) that estimated the intensity of the recombination process, tract length and mutational divergence between the true and candidate parental genomes (data not shown, details available from the authors). This approach additionally detected a larger number of smaller imports, suggesting that the heuristic approach was conservative.

Micro-evolution of P1 and P2 vaccine escape recombinants

The pattern of accumulation of sequence differences in the P1 and P2 vaccine escape recombinant lineages is highly heterogeneous: in P1 (42 isolates) the majority of isolates are identical or 1 nucleotide different from the ancestral sequence, across the filtered CustomSeq data (nominally ~12% of the genome) but branches leading to some isolates are much longer – CDC081 has 14 substitutions and CDC037 has 9 (Supplementary Figure 3 (A)). The short branches are consistent with mutation at a rate of $\sim 7 \times 10^{-7}$ $nt^{-1}yr^{-1}$ (although note that this data is not optimal for estimating substitution rates because of the uncertainty in sampling times), but longer branches (>2 substitutions) are not consistent with such low rates ($p << 0.01$ per branch for a Poisson model) or with recently published estimates ~2-fold higher [9] and furthermore in some cases show an obvious clustering of substitutions that would be extremely unlikely for randomly occurring mutations. Therefore we conclude that most if not all of the substitutions on such branches must have occurred through recombination, and note that the spacing of substitutions around the genome is only consistent with a multiple-transfer model of recombination involving at least 5 episodes of recombination (one per long branch). Approximately 32 substitutions occur on branches for which there is evidence of recombination at the $p = 0.01$ level and 23 on other branches, a proportion of $32/55 = 58\%$. These counts provide conservative estimates of the relative effect of recombination compared with mutation for at least two reasons: Firstly, branches with two substitutions are relatively unlikely to have arisen by a homogeneous mutation model but we have not counted them as recombinations. Secondly, even at the naively estimated rate of mutation some of the single-substitution branches are likely to have arisen by recombination. In addition, the filter we applied to the raw CustomSeq data is likely to have excluded clustered sites that would have looked like recombinations, although this will have been a small effect unless the average recombinational fragment size is small. Taking these considerations together, our data are not inconsistent with published estimates of the relative effects of recombination and mutation derived in two very different ways [9,10]: at least ~60% of substitutions we detect in the evolution of P1 are derived from recombination but we cannot exclude the possibility that

all (or nearly all) are. Whole-genome sequencing of the descendants of one or more well-defined recombination events will allow more precise resolution of this question.

The picture provided by the P2 lineage (8 isolates) (Supplementary Figure 3 (B)) is superficially different (there is no single dominant genotype) but is also consistent with a model in which multi-fragment recombination is the dominant mode of micro-evolution in *S. pneumoniae*: to explain the P2 data in terms of mutation would require a much higher estimated rate than for P1 and this seems unlikely. The lack of a dominant ancestral genotype in the data is broadly consistent with the observation that whereas the number of P1 individuals has increased dramatically since its emergence (a population expansion and a star-like phylogeny), no such expansion is evident for P2.

Supplementary References

1 Bandelt, H. J., Forster, P., Sykes, B. C. & Richards, M. B. Mitochondrial portraits of human populations using median networks. *Genetics* **141**, 743-753 (1995).

2 Pilishvili, T. *et al.* Sustained reductions in invasive pneumococcal disease in the era of conjugate vaccine. *J Infect Dis* **201**, 32-41 (2010).

3 Pai, R. *et al.* Postvaccine genetic structure of Streptococcus pneumoniae serotype 19A from children in the United States. *J Infect Dis* **192**, 1988-1995 (2005).

4 Brueggemann, A. B., Pai, R., Crook, D. W. & Beall, B. Vaccine escape recombinants emerge after pneumococcal vaccination in the United States. *PLoS Pathog* **3**, e168 (2007).

5 Cutler, D. J. *et al.* High-throughput variation detection and genotyping using microarrays. *Genome Res* **11**, 1913-1925 (2001).

6 Zwick, M. E. *et al.* Microarray-based resequencing of multiple Bacillus anthracis isolates. *Genome Biol* **6**, R10 (2005).

7 Tettelin, H. *et al.* Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. *Science* **293**, 498-506 (2001).

8 Darling, A. C., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**, 1394-1403 (2004).

9 Croucher, N. J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430-434 (2011).

10 Feil, E. J., Smith, J. M., Enright, M. C. & Spratt, B. G. Estimating recombinational parameters in Streptococcus pneumoniae from multilocus sequence typing data. *Genetics* **154**, 1439-1450 (2000).