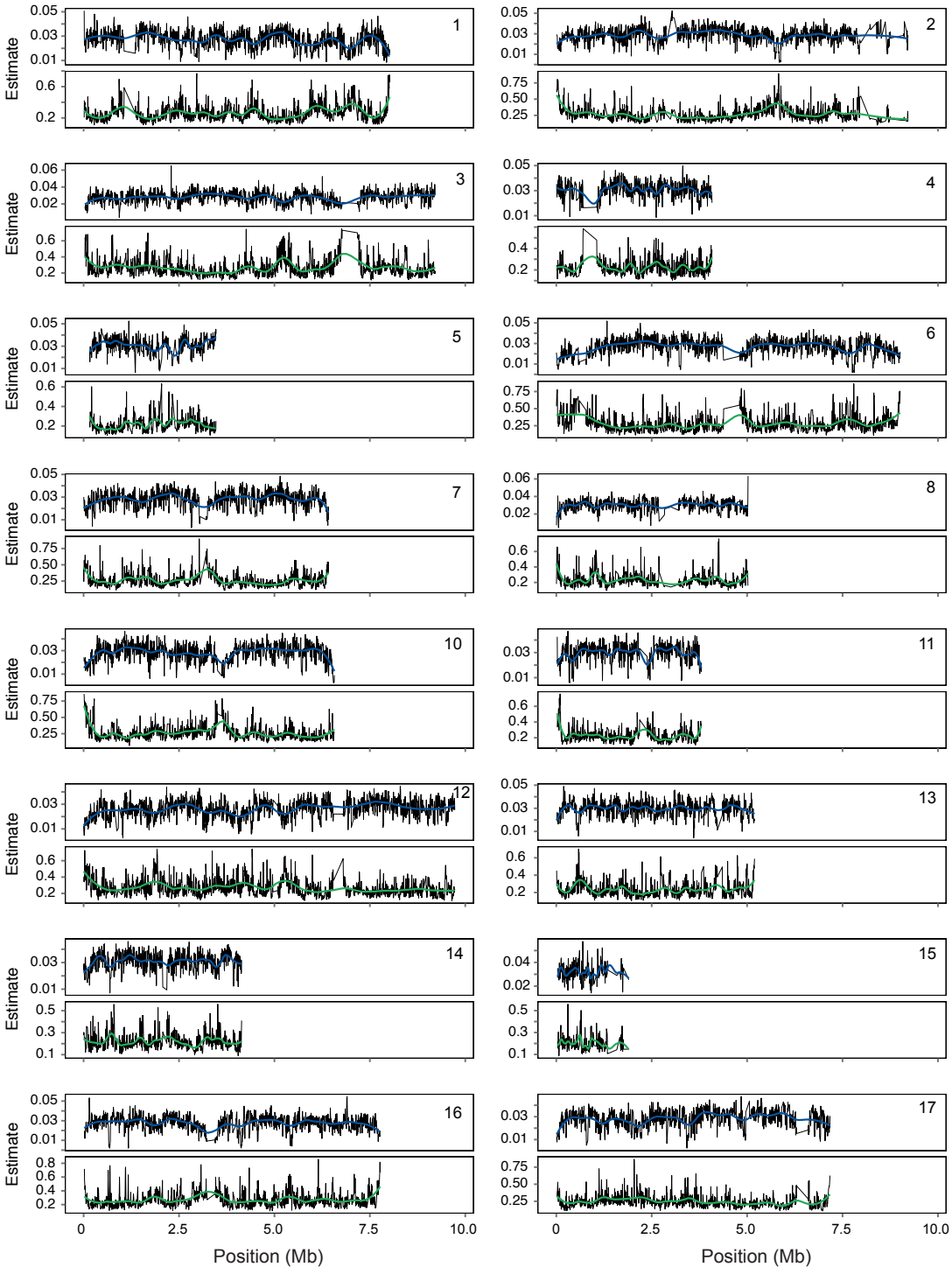
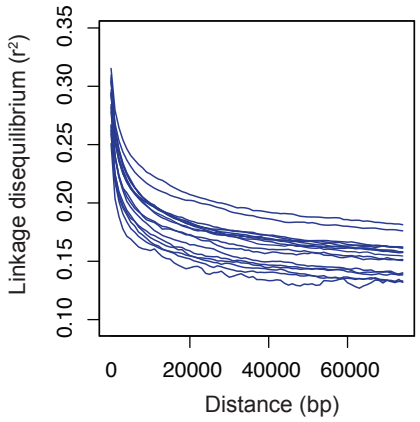
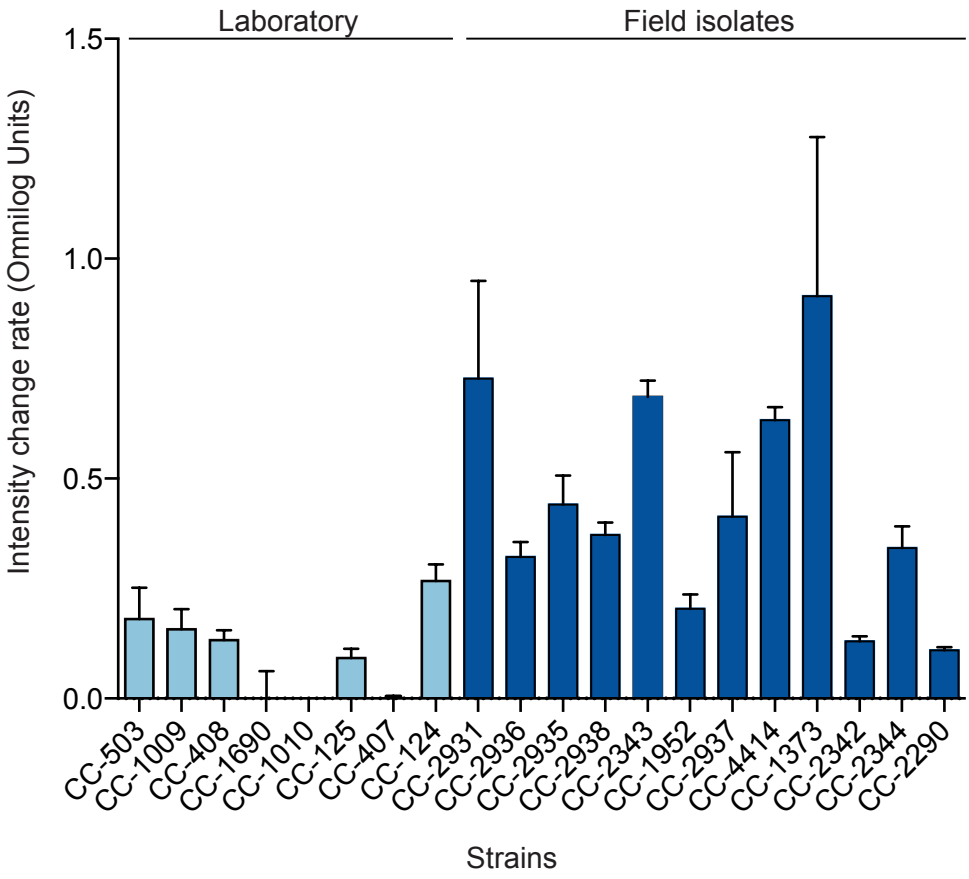


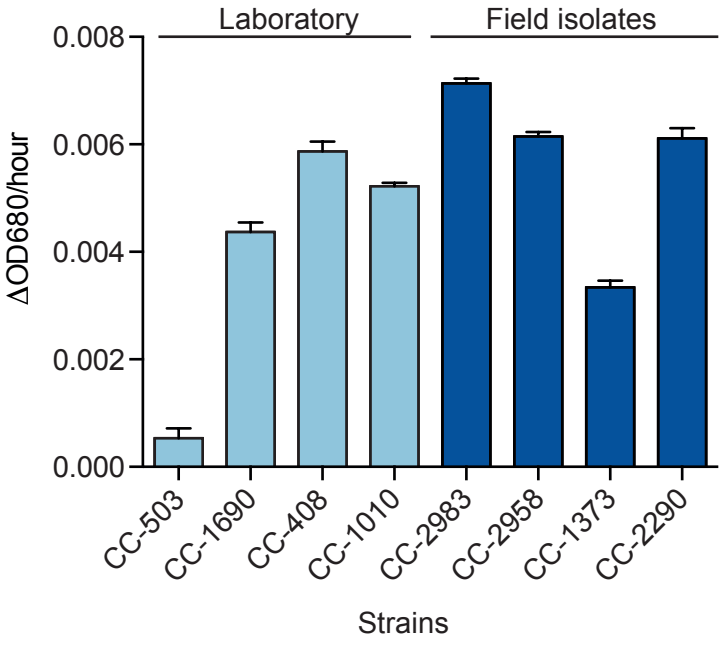
Supplemental Figure 1. Population genetic statistics on chromosomes of *C. reinhardtii*. Upper panels are nucleotide diversity (π), lower panels are Kelly's ZnS in non-overlapping windows of 5 kb. Trend lines were fit using Loess regression.



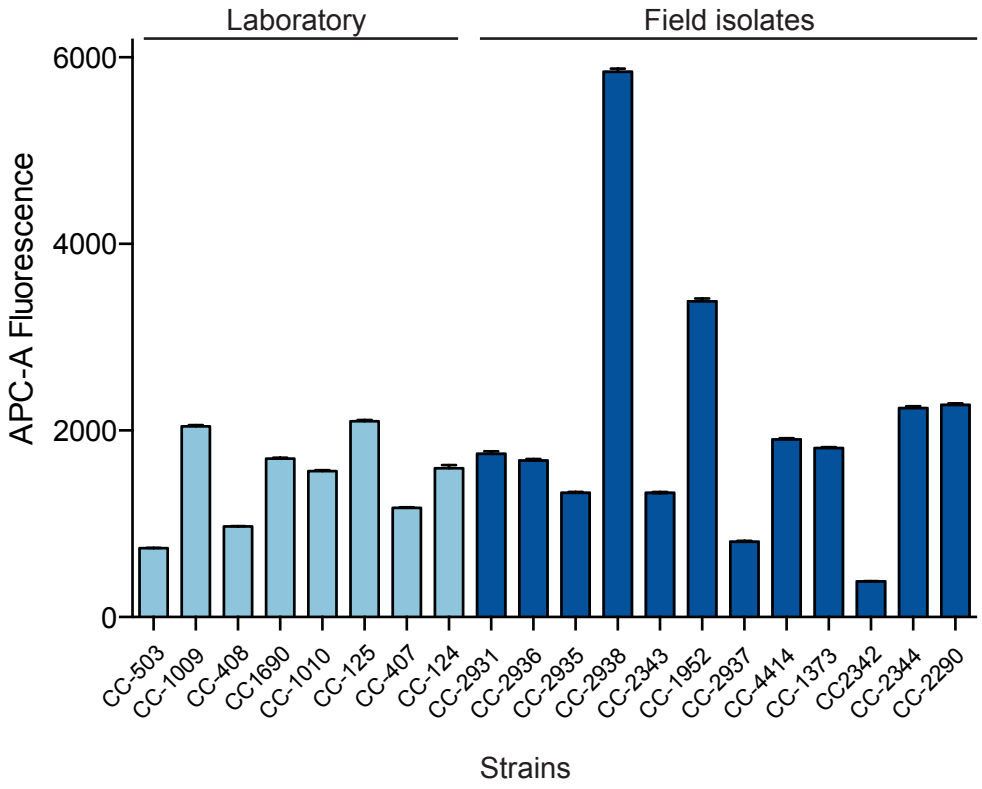
Supplemental Figure 2. Decay of linkage disequilibrium (LD) on 17 chromosomes of *C. reinhardtii*. LD is measured as the squared correlation coefficient (r^2) between SNP pairs.



Supplemental Figure 3. Growth rates of strains under heterotrophic conditions. The rates are measured as the amount of color intensity produced by the NADH reduction of a tetrazolium-based redox dye. Values are means + S.E.



Supplemental Figure 4. Growth rates for *C. reinhardtii* strains under phototrophic conditions. The rates are measured as change in absorbance at λ = 680 nm per hour for 8-12 days. Values are means + S.E.



Supplemental Figure 5. Chlorophyll content in strains of *C. reinhardtii.* Content was measured from Allophycocyanin (APC-A) fluorescence intensities. Values are means + S.E.

Supplemental Table 1. Summary of STRUCTURE (Pritchard et al. 2001) analyses.

K	Reps	Mean LnP(K)[a]	Stdev LnP(K)[b]	Ln'(K)[c]	|Ln''(K)|[c]	ΔK[c]
1	9	-55159.4222	25.1557 NA	NA	NA	NA
2	9	-52711.1667	172.9779	2448.255556	6100.744444	35.268925
3	9	-56363.6556	1087.372	-3652.488889	234411.4	215.576095
4	9	-294427.5444	289449.9449	-238063.8889	943048.7	3.258072
5	9	-1475540.133	1901995.603	-1181112.589	2105865.189	1.107187
6	9	-550787.5333	499001.422	924752.6	NA	NA

[a]Mean log-likelihood of K
[b]Standard deviation of the log likehood of K
[c]See definitions in Evanno et al. (2005)

Supplemental Table 2. Candidate major effect mutations in metabolic enzymes of *C. reinhardtii.*

Gene	Transcript	Name	Enzyme Name	E.C.	Mutation
Cre01.g012100	Cre01.g012100.t1.3	ARS4	-	3.1.6.-	deletion,nonsense
Cre01.g018900	Cre01.g018900.t1.2	NA	Ubiquitin--protein ligase	6.3.2.19	nonsense
Cre17.g710200	Cre17.g710200.t2.1	NA	"Indoleamine 2,3-dioxygenase."	1.13.11.52	nonsense
Cre02.g099850	Cre02.g099850.t1.3	PDC2	Pyruvate dehydrogenase (acetyl-transferring)	1.2.4.1	nonsense
Cre04.g223050	Cre04.g223050.t1.2	CAH2	Carbonate dehydratase	4.2.1.1	nonsense
Cre06.g308400	Cre06.g308400.t2.1	NA	"Phosphatidylinositol-3,4,5-trisphosphate 3-phosphatase"	3.1.3.67	nonsense
g6244	g6244.t1	STD1	Serine racemase	5.1.1.18	nonsense
Cre07.g321100	Cre07.g321100.t1.2	NA	Xylulokinase	2.7.1.17	nonsense
Cre08.g358900	Cre08.g358900.t1.2	NA	Hypoxanthine phosphoribosyltransferase	2.4.2.8	nonsense
g9116	g9116.t1	NA	-	2.4.1.-	deletion

Supplemental Table 3. Summary of *de novo* assembly results.

Samples	k-mer[a]	N50	Assembled sequence (bp)	Predicted genes	Genes with PFAM domain(s)	Genes with PFAM domain(s) after filtering
CC-125	37	671	495255	319	10	4
CC-1373	37	313	10198141	1323	56	23
CC-1690	37	509	1255235	501	26	12
CC-1952	41	303	12821542	1391	54	27
CC-2290	41	306	10071364	964	44	24
CC-2344	37	307	17213078	1865	66	16
CC-2343	37	330	15533271	1302	47	24
CC-2342	41	306	9918705	977	39	31
CC-2931	41	299	11361508	1366	63	26
CC-2935	41	303	6126924	768	41	20
CC-2936	41	322	6168226	858	107	61
CC-2937	37	313	8966638	1371	104	61
CC-2938	41	335	6126057	987	110	43

[a]k-mer length specified in Velvet (Zerbino and Birney 2008)

Supplemental Table 4. InterProScan5 predictions of PFAM domains for *de novo* assembled contigs. Gene model identifiers are default identifiers produced by Augustus. Identical gene model identifiers from different strains do not imply they are the same gene.

PFAM	Strain	Gene	Description
PF00004	CC-125	g28.t1	ATPase family associated with various cellular activities (AAA)
PF00069	CC-125	g1.t1	Protein kinase domain
PF00076	CC-125	g54.t1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00562	CC-125	g90.t1	RNA polymerase Rpb2, domain 6
PF02469	CC-125	g41.t1	Fasciclin domain
PF00005	CC-1373	g1066.t1	ABC transporter
PF00005	CC-1373	g202.t1	ABC transporter
PF00059	CC-1373	g10.t1	Lectin C-type domain
PF00069	CC-1373	g127.t1	Protein kinase domain
PF00069	CC-1373	g557.t1	Protein kinase domain
PF00069	CC-1373	g14.t1	Protein kinase domain
PF00069	CC-1373	g169.t1	Protein kinase domain
PF00069	CC-1373	g481.t1	Protein kinase domain
PF00076	CC-1373	g40.t1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00098	CC-1373	g200.t1	Zinc knuckle
PF00225	CC-1373	g524.t1	Kinesin motor domain
PF00225	CC-1373	g259.t1	Kinesin motor domain
PF00233	CC-1373	g369.t1	3'5'-cyclic nucleotide phosphodiesterase
PF00400	CC-1373	g208.t1	WD domain, G-beta repeat
PF00400	CC-1373	g1076.t1	WD domain, G-beta repeat
PF02469	CC-1373	g368.t1	Fasciclin domain
PF02469	CC-1373	g86.t1	Fasciclin domain
PF02785	CC-1373	g448.t1	Biotin carboxylase C-terminal domain
PF02786	CC-1373	g448.t1	Carbamoyl-phosphate synthase L chain, ATP binding domain
PF02892	CC-1373	g143.t1	BED zinc finger
PF02902	CC-1373	g323.t1	Ulp1 protease family, C-terminal catalytic domain
PF03372	CC-1373	g203.t1	Endonuclease/Exonuclease/phosphatase family
PF04811	CC-1373	g478.t1	Sec23/Sec24 trunk domain
PF04857	CC-1373	g571.t1	CAF1 family ribonuclease
PF05548	CC-1373	g388.t1	Gametolysin peptidase M11
PF00004	CC-1690	g149.t1	ATPase family associated with various cellular activities (AAA)
PF00069	CC-1690	g48.t1	Protein kinase domain
PF00076	CC-1690	g33.t1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00225	CC-1690	g268.t1	Kinesin motor domain
PF00400	CC-1690	g123.t1	WD domain, G-beta repeat
PF00400	CC-1690	g123.t1	WD domain, G-beta repeat

PF00562	CC-1690	g190.t1	RNA polymerase Rpb2, domain 6
PF01131	CC-1690	g11.t1	DNA topoisomerase
PF02469	CC-1690	g80.t1	Fasciclin domain
PF02889	CC-1690	g46.t1	Sec63 Brl domain
PF03016	CC-1690	g101.t1	Exostosin family
PF04811	CC-1690	g129.t1	Sec23/Sec24 trunk domain
PF04857	CC-1690	g132.t1	CAF1 family ribonuclease
PF00059	CC-1952	g52.t1	Lectin C-type domain
PF00059	CC-1952	g148.t1	Lectin C-type domain
PF00059	CC-1952	g148.t1	Lectin C-type domain
PF00059	CC-1952	g2.t1	Lectin C-type domain
PF00059	CC-1952	g2.t1	Lectin C-type domain
PF00059	CC-1952	g2.t1	Lectin C-type domain
PF00059	CC-1952	g2.t1	Lectin C-type domain
PF00059	CC-1952	g2.t1	Lectin C-type domain
PF00059	CC-1952	g229.t1	Lectin C-type domain
PF00069	CC-1952	g59.t1	Protein kinase domain
PF00069	CC-1952	g423.t1	Protein kinase domain
PF00076	CC-1952	g282.t1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00271	CC-1952	g571.t1	Helicase conserved C-terminal domain
PF00400	CC-1952	g47.t1	WD domain, G-beta repeat
PF00400	CC-1952	g47.t1	WD domain, G-beta repeat
PF00400	CC-1952	g47.t1	WD domain, G-beta repeat
PF00479	CC-1952	g187.t1	Glucose-6-phosphate dehydrogenase, NAD binding domain
PF00488	CC-1952	g794.t1	MutS domain V
PF00704	CC-1952	g229.t1	Glycosyl hydrolases family 18
PF00759	CC-1952	g344.t1	Glycosyl hydrolase family 9
PF01096	CC-1952	g437.t1	Transcription factor S-II (TFIIS)
PF02010	CC-1952	g38.t1	REJ domain
PF02373	CC-1952	g152.t1	JmjC domain, hydroxylase
PF02469	CC-1952	g77.t1	Fasciclin domain
PF02902	CC-1952	g319.t1	Ulp1 protease family, C-terminal catalytic domain
PF04434	CC-1952	g171.t1	SWIM zinc finger
PF04488	CC-1952	g169.t1	Glycosyltransferase sugar-binding region containing DXD motif
PF04857	CC-1952	g1069.t1	CAF1 family ribonuclease
PF06470	CC-1952	g301.t1	SMC proteins Flexible Hinge Domain
PF07393	CC-1952	g268.t1	Exocyst complex component Sec10
PF00059	CC-2290	g51.t1	Lectin C-type domain
PF00059	CC-2290	g51.t1	Lectin C-type domain
PF00059	CC-2290	g51.t1	Lectin C-type domain
PF00059	CC-2290	g51.t1	Lectin C-type domain
PF00059	CC-2290	g51.t1	Lectin C-type domain
PF00059	CC-2290	g210.t1	Lectin C-type domain
PF00059	CC-2290	g210.t1	Lectin C-type domain
PF00059	CC-2290	g512.t1	Lectin C-type domain
PF00059	CC-2290	g111.t1	Lectin C-type domain
PF00069	CC-2290	g195.t1	Protein kinase domain

PF00069	CC-2290	g84.t1	Protein kinase domain
PF00076	CC-2290	g4.t1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00271	CC-2290	g74.t1	Helicase conserved C-terminal domain
PF00400	CC-2290	g27.t1	WD domain, G-beta repeat
PF00400	CC-2290	g27.t1	WD domain, G-beta repeat
PF00400	CC-2290	g27.t1	WD domain, G-beta repeat
PF00530	CC-2290	g410.t1	Scavenger receptor cysteine-rich domain
PF00530	CC-2290	g110.t1	Scavenger receptor cysteine-rich domain
PF00704	CC-2290	g111.t1	Glycosyl hydrolases family 18
PF00759	CC-2290	g528.t1	Glycosyl hydrolase family 9
PF01096	CC-2290	g267.t1	Transcription factor S-II (TFIIS)
PF01476	CC-2290	g102.t1	LysM domain
PF02010	CC-2290	g321.t1	REJ domain
PF02373	CC-2290	g1.t1	JmjC domain, hydroxylase
PF02902	CC-2290	g300.t1	Ulp1 protease family, C-terminal catalytic domain
PF04434	CC-2290	g379.t1	SWIM zinc finger
PF04488	CC-2290	g6.t1	Glycosyltransferase sugar-binding region containing DXD motif
PF00059	CC-2342	g308.t1	Lectin C-type domain
PF00069	CC-2342	g29.t1	Protein kinase domain
PF00069	CC-2342	g65.t1	Protein kinase domain
PF00076	CC-2342	g52.t1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00112	CC-2342	g311.t1	Papain family cysteine protease
PF00202	CC-2342	g129.t1	Aminotransferase class-III
PF00754	CC-2342	g256.t1	F5/8 type C domain
PF00759	CC-2342	g536.t1	Glycosyl hydrolase family 9
PF01384	CC-2342	g145.t1	Phosphate transporter family
PF01384	CC-2342	g88.t1	Phosphate transporter family
PF01476	CC-2342	g14.t1	LysM domain
PF01476	CC-2342	g14.t1	LysM domain
PF02010	CC-2342	g296.t1	REJ domain
PF02892	CC-2342	g237.t1	BED zinc finger
PF02902	CC-2342	g269.t1	Ulp1 protease family, C-terminal catalytic domain
PF02902	CC-2342	g194.t1	Ulp1 protease family, C-terminal catalytic domain
PF03184	CC-2342	g103.t1	DDE superfamily endonuclease
PF00059	CC-2343	g145.t1	Lectin C-type domain
PF00059	CC-2343	g145.t1	Lectin C-type domain
PF00059	CC-2343	g22.t1	Lectin C-type domain
PF00059	CC-2343	g22.t1	Lectin C-type domain
PF00059	CC-2343	g64.t1	Lectin C-type domain
PF00059	CC-2343	g64.t1	Lectin C-type domain
PF00059	CC-2343	g64.t1	Lectin C-type domain
PF00059	CC-2343	g274.t1	Lectin C-type domain
PF00069	CC-2343	g83.t1	Protein kinase domain

PF00069	CC-2343	g17.t1	Protein kinase domain
PF00069	CC-2343	g77.t1	Protein kinase domain
PF00069	CC-2343	g450.t1	Protein kinase domain
PF00069	CC-2343	g1248.t1	Protein kinase domain
PF00076	CC-2343	g69.t1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00211	CC-2343	g20.t1	Adenylate and Guanylate cyclase catalytic domain
PF00246	CC-2343	g437.t1	Zinc carboxypeptidase
PF00350	CC-2343	g638.t1	Dynamin family
PF00530	CC-2343	g480.t1	Scavenger receptor cysteine-rich domain
PF00530	CC-2343	g323.t1	Scavenger receptor cysteine-rich domain
PF00704	CC-2343	g274.t1	Glycosyl hydrolases family 18
PF01094	CC-2343	g82.t1	Receptor family ligand binding region
PF01753	CC-2343	g199.t1	MYND finger
PF02010	CC-2343	g293.t1	REJ domain
PF02469	CC-2343	g85.t1	Fasciclin domain
PF02469	CC-2343	g208.t1	Fasciclin domain
PF00004	CC-2344	g38.t1	ATPase family associated with various cellular activities (AAA)
PF00059	CC-2344	g81.t1	Lectin C-type domain
PF00059	CC-2344	g81.t1	Lectin C-type domain
PF00059	CC-2344	g81.t1	Lectin C-type domain
PF00059	CC-2344	g43.t1	Lectin C-type domain
PF00059	CC-2344	g43.t1	Lectin C-type domain
PF00059	CC-2344	g129.t1	Lectin C-type domain
PF00059	CC-2344	g129.t1	Lectin C-type domain
PF00059	CC-2344	g92.t1	Lectin C-type domain
PF00069	CC-2344	g772.t1	Protein kinase domain
PF00069	CC-2344	g297.t1	Protein kinase domain
PF00069	CC-2344	g69.t1	Protein kinase domain
PF00069	CC-2344	g484.t1	Protein kinase domain
PF00076	CC-2344	g126.t1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00176	CC-2344	g858.t1	SNF2 family N-terminal domain
PF00225	CC-2344	g929.t1	Kinesin motor domain
PF00488	CC-2344	g195.t1	MutS domain V
PF00530	CC-2344	g285.t1	Scavenger receptor cysteine-rich domain
PF00530	CC-2344	g285.t1	Scavenger receptor cysteine-rich domain
PF00562	CC-2344	g889.t1	RNA polymerase Rpb2, domain 6
PF00628	CC-2344	g578.t1	PHD-finger
PF00704	CC-2344	g92.t1	Glycosyl hydrolases family 18
PF00999	CC-2344	g1003.t1	Sodium/hydrogen exchanger family
PF01624	CC-2344	g274.t1	MutS domain I
PF02010	CC-2344	g163.t1	REJ domain
PF02469	CC-2344	g20.t1	Fasciclin domain
PF02902	CC-2344	g576.t1	Ulp1 protease family, C-terminal

			catalytic domain
PF03372	CC-2344	g683.t1	Endonuclease/Exonuclease/phosphatase family
PF04479	CC-2344	g201.t1	RTA1 like protein
PF04857	CC-2344	g811.t1	CAF1 family ribonuclease
PF05190	CC-2344	g195.t1	MutS family domain IV
PF06985	CC-2344	g258.t1	Heterokaryon incompatibility protein (HET)
PF00023	CC-2931	g197.t1	Ankyrin repeat
PF00059	CC-2931	g221.t1	Lectin C-type domain
PF00059	CC-2931	g221.t1	Lectin C-type domain
PF00059	CC-2931	g22.t1	Lectin C-type domain
PF00059	CC-2931	g22.t1	Lectin C-type domain
PF00059	CC-2931	g165.t1	Lectin C-type domain
PF00059	CC-2931	g165.t1	Lectin C-type domain
PF00059	CC-2931	g97.t1	Lectin C-type domain
PF00069	CC-2931	g62.t1	Protein kinase domain
PF00076	CC-2931	g142.t1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00112	CC-2931	g55.t1	Papain family cysteine protease
PF00246	CC-2931	g384.t1	Zinc carboxypeptidase
PF00350	CC-2931	g428.t1	Dynamin family
PF00530	CC-2931	g58.t1	Scavenger receptor cysteine-rich domain
PF00530	CC-2931	g47.t1	Scavenger receptor cysteine-rich domain
PF00530	CC-2931	g47.t1	Scavenger receptor cysteine-rich domain
PF00704	CC-2931	g97.t1	Glycosyl hydrolases family 18
PF00754	CC-2931	g25.t1	F5/8 type C domain
PF01384	CC-2931	g342.t1	Phosphate transporter family
PF01476	CC-2931	g39.t1	LysM domain
PF02010	CC-2931	g67.t1	REJ domain
PF02338	CC-2931	g156.t1	OTU-like cysteine protease
PF02902	CC-2931	g200.t1	Ulp1 protease family, C-terminal catalytic domain
PF02902	CC-2931	g27.t1	Ulp1 protease family, C-terminal catalytic domain
PF03184	CC-2931	g572.t1	DDE superfamily endonuclease
PF03372	CC-2931	g185.t1	Endonuclease/Exonuclease/phosphatase family
PF05970	CC-2931	g464.t1	PIF1-like helicase
PF00023	CC-2935	g438.t1	Ankyrin repeat
PF00059	CC-2935	g31.t1	Lectin C-type domain
PF00059	CC-2935	g31.t1	Lectin C-type domain
PF00059	CC-2935	g251.t1	Lectin C-type domain
PF00059	CC-2935	g8.t1	Lectin C-type domain
PF00059	CC-2935	g115.t1	Lectin C-type domain
PF00059	CC-2935	g115.t1	Lectin C-type domain
PF00059	CC-2935	g78.t1	Lectin C-type domain
PF00059	CC-2935	g11.t1	Lectin C-type domain

PF00069	CC-2935	g389.t1	Protein kinase domain
PF00069	CC-2935	g363.t1	Protein kinase domain
PF00076	CC-2935	g434.t1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00112	CC-2935	g472.t1	Papain family cysteine protease
PF00530	CC-2935	g87.t1	Scavenger receptor cysteine-rich domain
PF00530	CC-2935	g368.t1	Scavenger receptor cysteine-rich domain
PF00704	CC-2935	g11.t1	Glycosyl hydrolases family 18
PF00754	CC-2935	g46.t1	F5/8 type C domain
PF03184	CC-2935	g42.t1	DDE superfamily endonuclease
PF03184	CC-2935	g328.t1	DDE superfamily endonuclease
PF05729	CC-2935	g135.t1	NACHT domain
PF05970	CC-2935	g171.t1	PIF1-like helicase
PF00005	CC-2936	g41.t1	ABC transporter
PF00023	CC-2936	g99.t1	Ankyrin repeat
PF00069	CC-2936	g197.t1	Protein kinase domain
PF00069	CC-2936	g268.t1	Protein kinase domain
PF00069	CC-2936	g663.t1	Protein kinase domain
PF00069	CC-2936	g564.t1	Protein kinase domain
PF00076	CC-2936	g575.t1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00136	CC-2936	g57.t1	DNA polymerase family B
PF00136	CC-2936	g57.t1	DNA polymerase family B
PF00136	CC-2936	g57.t1	DNA polymerase family B
PF00176	CC-2936	g72.t1	SNF2 family N-terminal domain
PF00176	CC-2936	g129.t1	SNF2 family N-terminal domain
PF00204	CC-2936	g13.t1	DNA gyrase B
PF00271	CC-2936	g129.t1	Helicase conserved C-terminal domain
PF00317	CC-2936	g63.t1	Ribonucleotide reductase, all-alpha domain
PF00350	CC-2936	g523.t1	Dynamin family
PF00385	CC-2936	g141.t1	Chromo (CHRromatin Organisation MOdifier) domain
PF00521	CC-2936	g13.t1	DNA gyrase/topoisomerase IV, subunit A
PF00562	CC-2936	g101.t1	RNA polymerase Rpb2, domain 6
PF00562	CC-2936	g82.t1	RNA polymerase Rpb2, domain 6
PF00562	CC-2936	g430.t1	RNA polymerase Rpb2, domain 6
PF00580	CC-2936	g306.t1	UvrD/REP helicase N-terminal domain
PF00580	CC-2936	g215.t1	UvrD/REP helicase N-terminal domain
PF00623	CC-2936	g43.t1	RNA polymerase Rpb1, domain 2
PF00636	CC-2936	g82.t1	Ribonuclease III domain
PF00656	CC-2936	g102.t1	Caspase domain
PF00850	CC-2936	g500.t1	Histone deacetylase domain
PF01096	CC-2936	g137.t1	Transcription factor S-II (TFIIS)
PF01193	CC-2936	g42.t1	RNA polymerase Rpb3/Rpb11 dimerisation domain
PF01331	CC-2936	g132.t1	mRNA capping enzyme, catalytic domain

PF01652	CC-2936	g11.t1	Eukaryotic initiation factor 4E
PF01712	CC-2936	g110.t1	Deoxynucleoside kinase
PF01734	CC-2936	g41.t1	Patatin-like phospholipase
PF01751	CC-2936	g13.t1	Toprim domain
PF01753	CC-2936	g515.t1	MYND finger
PF02373	CC-2936	g69.t1	JmjC domain, hydroxylase
PF02469	CC-2936	g127.t1	Fasciclin domain
PF02518	CC-2936	g13.t1	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase
PF02739	CC-2936	g172.t1	5'-3' exonuclease, N-terminal resolvase-like domain
PF02867	CC-2936	g63.t1	Ribonucleotide reductase, barrel domain
PF02902	CC-2936	g259.t1	Ulp1 protease family, C-terminal catalytic domain
PF02940	CC-2936	g132.t1	mRNA capping enzyme, beta chain
PF03104	CC-2936	g57.t1	DNA polymerase family B, exonuclease domain
PF03104	CC-2936	g57.t1	DNA polymerase family B, exonuclease domain
PF03109	CC-2936	g140.t1	ABC1 family
PF03159	CC-2936	g136.t1	XRN 5'-3' exonuclease N-terminus
PF03184	CC-2936	g381.t1	DDE superfamily endonuclease
PF03291	CC-2936	g132.t1	mRNA capping enzyme
PF03291	CC-2936	g132.t1	mRNA capping enzyme
PF03372	CC-2936	g149.t1	Endonuclease/Exonuclease/phosphatase family
PF04560	CC-2936	g82.t1	RNA polymerase Rpb2, domain 7
PF04560	CC-2936	g430.t1	RNA polymerase Rpb2, domain 7
PF04563	CC-2936	g101.t1	RNA polymerase beta subunit
PF04563	CC-2936	g82.t1	RNA polymerase beta subunit
PF04565	CC-2936	g101.t1	RNA polymerase Rpb2, domain 3
PF04565	CC-2936	g101.t1	RNA polymerase Rpb2, domain 3
PF04565	CC-2936	g82.t1	RNA polymerase Rpb2, domain 3
PF04566	CC-2936	g101.t1	RNA polymerase Rpb2, domain 4
PF04566	CC-2936	g82.t1	RNA polymerase Rpb2, domain 4
PF04567	CC-2936	g101.t1	RNA polymerase Rpb2, domain 5
PF04567	CC-2936	g82.t1	RNA polymerase Rpb2, domain 5
PF04983	CC-2936	g43.t1	RNA polymerase Rpb1, domain 3
PF04997	CC-2936	g43.t1	RNA polymerase Rpb1, domain 1
PF04998	CC-2936	g43.t1	RNA polymerase Rpb1, domain 5
PF04998	CC-2936	g43.t1	RNA polymerase Rpb1, domain 5
PF04998	CC-2936	g43.t1	RNA polymerase Rpb1, domain 5
PF05000	CC-2936	g43.t1	RNA polymerase Rpb1, domain 4
PF00004	CC-2937	g105.t1	ATPase family associated with various cellular activities (AAA)
PF00005	CC-2937	g7.t1	ABC transporter
PF00069	CC-2937	g76.t1	Protein kinase domain
PF00069	CC-2937	g58.t1	Protein kinase domain
PF00076	CC-2937	g337.t1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)

PF00136	CC-2937	g419.t1	DNA polymerase family B
PF00136	CC-2937	g187.t1	DNA polymerase family B
PF00136	CC-2937	g187.t1	DNA polymerase family B
PF00176	CC-2937	g35.t1	SNF2 family N-terminal domain
PF00176	CC-2937	g48.t1	SNF2 family N-terminal domain
PF00204	CC-2937	g97.t1	DNA gyrase B
PF00271	CC-2937	g35.t1	Helicase conserved C-terminal domain
PF00271	CC-2937	g701.t1	Helicase conserved C-terminal domain
PF00317	CC-2937	g168.t1	Ribonucleotide reductase, all-alpha domain
PF00400	CC-2937	g205.t1	WD domain, G-beta repeat
PF00488	CC-2937	g89.t1	MutS domain V
PF00521	CC-2937	g97.t1	DNA gyrase/topoisomerase IV, subunit A
PF00562	CC-2937	g930.t1	RNA polymerase Rpb2, domain 6
PF00562	CC-2937	g118.t1	RNA polymerase Rpb2, domain 6
PF00562	CC-2937	g317.t1	RNA polymerase Rpb2, domain 6
PF00580	CC-2937	g47.t1	UvrD/REP helicase N-terminal domain
PF00623	CC-2937	g9.t1	RNA polymerase Rpb1, domain 2
PF00656	CC-2937	g316.t1	Caspase domain
PF00754	CC-2937	g792.t1	F5/8 type C domain
PF00850	CC-2937	g177.t1	Histone deacetylase domain
PF01193	CC-2937	g8.t1	RNA polymerase Rpb3/Rpb11 dimerisation domain
PF01331	CC-2937	g366.t1	mRNA capping enzyme, catalytic domain
PF01384	CC-2937	g329.t1	Phosphate transporter family
PF01384	CC-2937	g174.t1	Phosphate transporter family
PF01652	CC-2937	g95.t1	Eukaryotic initiation factor 4E
PF01734	CC-2937	g7.t1	Patatin-like phospholipase
PF01751	CC-2937	g97.t1	Toprim domain
PF01753	CC-2937	g270.t1	MYND finger
PF02037	CC-2937	g975.t1	SAP domain
PF02373	CC-2937	g171.t1	JmjC domain, hydroxylase
PF02469	CC-2937	g185.t1	Fasciclin domain
PF02518	CC-2937	g97.t1	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase
PF02739	CC-2937	g420.t1	5'-3' exonuclease, N-terminal resolvase-like domain
PF02867	CC-2937	g168.t1	Ribonucleotide reductase, barrel domain
PF02902	CC-2937	g134.t1	Ulp1 protease family, C-terminal catalytic domain
PF02902	CC-2937	g122.t1	Ulp1 protease family, C-terminal catalytic domain
PF02940	CC-2937	g366.t1	mRNA capping enzyme, beta chain
PF03104	CC-2937	g105.t1	DNA polymerase family B, exonuclease domain
PF03104	CC-2937	g105.t1	DNA polymerase family B, exonuclease domain
PF03109	CC-2937	g279.t1	ABC1 family

PF03159	CC-2937	g264.t1	XRN 5'-3' exonuclease N-terminus
PF03184	CC-2937	g369.t1	DDE superfamily endonuclease
PF03291	CC-2937	g37.t1	mRNA capping enzyme
PF03291	CC-2937	g37.t1	mRNA capping enzyme
PF03372	CC-2937	g50.t1	Endonuclease/Exonuclease/phosphatase family
PF04479	CC-2937	g188.t1	RTA1 like protein
PF04560	CC-2937	g118.t1	RNA polymerase Rpb2, domain 7
PF04563	CC-2937	g118.t1	RNA polymerase beta subunit
PF04563	CC-2937	g317.t1	RNA polymerase beta subunit
PF04565	CC-2937	g118.t1	RNA polymerase Rpb2, domain 3
PF04565	CC-2937	g317.t1	RNA polymerase Rpb2, domain 3
PF04565	CC-2937	g317.t1	RNA polymerase Rpb2, domain 3
PF04566	CC-2937	g118.t1	RNA polymerase Rpb2, domain 4
PF04566	CC-2937	g317.t1	RNA polymerase Rpb2, domain 4
PF04567	CC-2937	g118.t1	RNA polymerase Rpb2, domain 5
PF04567	CC-2937	g317.t1	RNA polymerase Rpb2, domain 5
PF04983	CC-2937	g9.t1	RNA polymerase Rpb1, domain 3
PF04997	CC-2937	g9.t1	RNA polymerase Rpb1, domain 1
PF04998	CC-2937	g9.t1	RNA polymerase Rpb1, domain 5
PF04998	CC-2937	g9.t1	RNA polymerase Rpb1, domain 5
PF04998	CC-2937	g9.t1	RNA polymerase Rpb1, domain 5
PF05000	CC-2937	g9.t1	RNA polymerase Rpb1, domain 4
PF05970	CC-2937	g404.t1	PIF1-like helicase
PF00004	CC-2938	g87.t1	ATPase family associated with various cellular activities (AAA)
PF00011	CC-2938	g27.t1	Hsp20/alpha crystallin family
PF00012	CC-2938	g63.t1	Hsp70 protein
PF00023	CC-2938	g491.t1	Ankyrin repeat
PF00069	CC-2938	g677.t1	Protein kinase domain
PF00069	CC-2938	g90.t1	Protein kinase domain
PF00069	CC-2938	g488.t1	Protein kinase domain
PF00069	CC-2938	g4.t1	Protein kinase domain
PF00069	CC-2938	g261.t1	Protein kinase domain
PF00069	CC-2938	g421.t1	Protein kinase domain
PF00075	CC-2938	g99.t1	RNase H
PF00075	CC-2938	g212.t1	RNase H
PF00076	CC-2938	g507.t1	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
PF00125	CC-2938	g376.t1	Core histone H2A/H2B/H3/H4
PF00125	CC-2938	g56.t1	Core histone H2A/H2B/H3/H4
PF00183	CC-2938	g83.t1	Hsp90 protein
PF00226	CC-2938	g13.t1	DnaJ domain
PF00270	CC-2938	g149.t1	DEAD/DEAH box helicase
PF00271	CC-2938	g149.t1	Helicase conserved C-terminal domain
PF00303	CC-2938	g17.t1	Thymidylate synthase
PF00317	CC-2938	g68.t1	Ribonucleotide reductase, all-alpha domain
PF00382	CC-2938	g89.t1	Transcription factor TFIIB repeat
PF00498	CC-2938	g330.t1	FHA domain
PF00580	CC-2938	g154.t1	UvrD/REP helicase N-terminal domain

PF00642	CC-2938	g163.t1	Zinc finger C-x8-C-x5-C-x3-H type (and similar)
PF00651	CC-2938	g630.t1	BTB/POZ domain
PF00656	CC-2938	g43.t1	Caspase domain
PF00692	CC-2938	g162.t1	dUTPase
PF00733	CC-2938	g8.t1	Asparagine synthase
PF00752	CC-2938	g169.t1	XPG N-terminal domain
PF00773	CC-2938	g147.t1	RNB domain
PF00773	CC-2938	g147.t1	RNB domain
PF00782	CC-2938	g305.t1	Dual specificity phosphatase, catalytic domain
PF00867	CC-2938	g169.t1	XPG I-region
PF00929	CC-2938	g425.t1	Exonuclease
PF01068	CC-2938	g60.t1	ATP dependent DNA ligase domain
PF01233	CC-2938	g273.t1	Myristoyl-CoA:protein N-myristoyltransferase, N-terminal domain
PF01453	CC-2938	g69.t1	D-mannose binding lectin
PF01453	CC-2938	g69.t1	D-mannose binding lectin
PF01556	CC-2938	g13.t1	DnaJ C terminal domain
PF01636	CC-2938	g90.t1	Phosphotransferase enzyme family
PF02373	CC-2938	g105.t1	JmjC domain, hydroxylase
PF02469	CC-2938	g94.t1	Fasciclin domain
PF02799	CC-2938	g273.t1	Myristoyl-CoA:protein N-myristoyltransferase, C-terminal domain
PF02867	CC-2938	g68.t1	Ribonucleotide reductase, barrel domain
PF02902	CC-2938	g255.t1	Ulp1 protease family, C-terminal catalytic domain
PF03291	CC-2938	g16.t1	mRNA capping enzyme
PF03372	CC-2938	g72.t1	Endonuclease/Exonuclease/phosphatase family
PF04479	CC-2938	g374.t1	RTA1 like protein
PF04675	CC-2938	g60.t1	DNA ligase N terminus
PF04679	CC-2938	g60.t1	ATP dependent DNA ligase C terminal region
PF05050	CC-2938	g191.t1	Methyltransferase FkbM domain
PF05548	CC-2938	g306.t1	Gametolysin peptidase M11
PF05548	CC-2938	g128.t1	Gametolysin peptidase M11
PF05548	CC-2938	g96.t1	Gametolysin peptidase M11
PF05970	CC-2938	g343.t1	PIF1-like helicase

Supplemental Table 5. Genome segments in laboratory strains that are not identical-by-descent (IBD) with CC-503. The boundaries of the segments were identified by manual inspection of non-reference SNP counts in non-overlapping windows of 25 kb. The reported intervals are inclusive of the boundary but may include a small number of bases that are IBD with CC-503 owing to a lack of precision using the window-based approach.

Chr	Start	End	Strain(s)
1	1725000	3225000	(CC-1009,CC 408)
1	3500000	4275000	(CC-1009,CC-408)
1	4300000	6025000	(CC-1009,CC-408)
1	6050000	6350000	(CC-1009,CC-408)
2	4075000	4875000	(CC-1010,CC-407,CC-1690)
2	6150000	6950000	(CC-1010,CC-407,CC-1690)
3	8400000	8650000	(CC-124,CC-1009,CC-408)
3	8675000	9025000	(CC-124,CC-1009,CC-408)
4	150000	250000	(CC-1010,CC-407,CC-1690)
6	0	625000	(CC-124,CC-1009,CC-408)
6	775000	1100000	(CC-124,CC-1009,CC-408)
6	1250000	1975000	(CC-124,CC-1009,CC-408)
8	125000	150000	(CC-1010,CC-407,CC-1690)
8	225000	725000	(CC-1010,CC-407,CC-1690)
9	0	325000	(CC-1009,CC-408)
9	350000	725000	(CC-1009,CC-408)
9	750000	2125000	(CC-1009,CC-408)
9	2275000	2400000	(CC-1009,CC-408)
9	2925000	3100000	(CC-1010,CC-407,CC-1690)
10	0	650000	(CC-1010,CC-407,CC-1690)
10	5875000	6150000	(CC-1009,CC-408)
10	6200000	6575000	(CC-1009,CC-408)
11	0	1025000	(CC-1009,CC-408)
11	1200000	1300000	(CC-1009,CC-408)
11	1275000	1350000	(CC-1010,CC-407,CC-1690)
11	1375000	2225000	(CC-1010,CC-407,CC-1690)
12	0	450000	(CC-1009,CC-408)
12	500000	825000	(CC-1009,CC-408)
12	1000000	1850000	(CC-1009,CC-408)
12	1925000	2275000	(CC-1009,CC-408)
12	2300000	2325000	(CC-1009,CC-408)
12	2350000	2475000	(CC-1009,CC-408)
12	7225000	7825000	(CC-1009,CC-408)
12	7925000	8025000	(CC-1009,CC-408)
12	8050000	8675000	(CC-1009,CC-408)
12	8950000	9730733	(CC-124,CC-1009,CC-408)
15	850000	1300000	(CC-1010,CC-407,CC-1690)
16	0	450000	(CC-1009,CC-408)
16	475000	750000	(CC-1009,CC-408)
16	850000	975000	(CC-1010,CC-407,CC-1690,CC-124)
16	1425000	1950000	(CC-1010,CC-407,CC-

			1690,CC-124)
16	6325000	7783580	(CC-124,CC-1009,CC-408)
17	275000	300000	(CC-124,CC-1009,CC-408)
17	325000	375000	(CC-124,CC-1009,CC-408)
17	400000	1450000	(CC-124,CC-1009,CC-408)
17	2125000	2800000	(CC-1009,CC-408)
17	2900000	3350000	(CC-1009,CC-408)
17	3800000	4700000	(CC-1009,CC-408)
17	5050000	5525000	(CC-1009,CC-408)
17	5525000	6075000	(CC-1009,CC-408)

Supplemental Table 6. Chromosomal breakpoints of large copy number gains in strains of *C. reinhardtii* relative to CC-503. Breakpoints were identified by manually inspecting normalized coverage depth in 499 bp windows.

Strain	Chr	Start	End
CC-407	1	4318500	4725000
CC-407	6	4273000	4364500
CC-407	6	4370500	4447000
CC-407	6	4756000	4825000
CC-407	12	5720500	5863500
CC-1010	13	4141500	4227500
CC-1010	13	4349500	4403500
CC-1010	13	4487000	4537000
CC-1010	17	5998000	6333000
CC-2290	9	5286500	5319500
CC-2290	9	5342500	5353000
CC-2290	9	5375000	5476000
CC-2290	17	4278500	end
CC-2343	9	5286500	5467000
CC-2935	8	1000500	1078000
CC-2935	13	4229000	4598500
CC-1373	9	5306000	5519500
CC-1373	8	4556000	end
CC-1373	14	2105500	2181500
CC-2344	9	5283000	5470000
CC-2931	9	5283000	5470000
CC-2937	9	5306500	5462000
CC-1952	8	4376500	4472000
CC-1952	8	4699500	4852500
CC-1952	8	4891000	end
CC-1952	9	5283500	5476000
CC-1690	13	4141500	4227500
CC-1690	13	4349500	4403500
CC-1690	13	4487000	4537000
CC-2342	13	2174500	end

Supplemental Table 7. Summary of SNP-filtering protocol including cut-off thresholds.

Filter	**Condition**[a]
Low depth	DP < 385
High depth	DP > 2285
QUAL[b]	QUAL < 50
QD[b]	QD < 3
FS[b]	FS > 775
MQranksum[b]	MQranksum < -31.50
ReadPosRankSum[b]	ReadPosRankSum < -3.5
SNPs with heterozygotes genotype(s)	All
Low complexity regions	All
SNPs with a non-reference genotype in CC-503	All

[a]SNPs meeting the indicated condition were removed from the analysis
[b]See https://www.broadinstitute.org/gatk/ for definition

Supplemental Table 8. PCR and Sanger-based sequencing validation of randomly selected nonsense SNPs.

Chr	Position	Transcript	AA Change[a]	REF[b]-Illumina	REF-Sanger	ALT[c]-Illumina	ALT-Sanger	REF-Strain	ALT-Strain
1	2209754	Cre01.g012100.t1.3	Y267*	T	T	G	G	125	1373
1	2210377	Cre01.g012100.t1.3	Y321*	C	-	A	A	-	1373
1	2215883	Cre01.g012126.t1.2	Y48*	C	C	A	A	125, 2937	2935
1	3533275	g523.t1	W14*	G	G	A	A	125	2931
1	7078063	g1132.t1	Q31*	C	C	T	T	2931	1373
2	4421439	Cre02.g099850.t1.3	W30*	G	G	A	A	125, 2290	2937
2	4421646	Cre02.g099850.t1.3	L44*	T	T	A	A	125, 2290	2937
3	1409312	Cre03.g151050.t1.3	Q86*	C	C	T	T	2838	2931
3	7198504	Cre03.g207918.t1.2	Y294*	T	T	A	A	125	2343
4	15343	g4492.t1	W247*	G	G	A	A	2343	2290
4	2175298	Cre04.g219796.t1.2	W206*	G	G	A	A	2935	2342
4	2304507	g4809.t1	W85*	C	C	T	T	125	2290
5	313924	Cre05.g233400.t1.3	E5*	G	G	T	T	125	2290
6	1150879	g5779.t1	Y32*	G	G	C	C	125	2935
6	1651003	g5865.t1	Q19*	G	G	A	A	125	2938
6	2349518	g6002.t1	K81*	T	T	A	A	125	2931
6	3486202	g6244.t1	R19*	C	C	T	T	1373, 2935	2344
7	2617704	g7686.t1	Q64*	C	C	T	T	125	2931
10	210521	Cre10.g419200.t1.3	S59*	C	C	A	A	2931, 2343	2935
10	713044	Cre10.g422550.t1.2	E746*	C	C	A	A	125	2935
10	5378206	Cre10.g458350.t3.1	Q552*	C	C	T	T	125	2343
11	7934	g11450.t1	S91*	G	G	T	G	125	2342, 1952, 2290
11	8163	g11450.t1	R15*	G	G	A	G	2931	1952, 2290, 2342, 2343, 2344
11	626316	g11543.t1	W13*	C	C	T	T	125	2935

11	3601821	Cre11.g481950.t1.3	Q154*	C	C	T	T	125	2935
12	1287125	g12203.t1	S50*	G	-	C	C	-	2831
12	9597928	g13805.t1	R21*	C	C	T	T	125, 2342	2290
13	451796	Cre13.g564840.t1.3	Y319*	G	G	T	T	125	2290
13	1764055	Cre13.g574800.t1.2	S182*	C	C	A	A	125	2290
15	760377	Cre15.g640152.t1.2	Y830*	G	G	T	T	124	2344, 2342
15	1762528	Cre15.g643500.t1.3	W350*	G	G	A	A	2931, 2290	2343
16	934535	Cre16.g648500.t1.2	Q55*	G	G	A	A	125	2290
16	4136903	Cre16.g669150.t1.2	Y66*	G	G	C	C	2342	2344

[a]Amino acid change
[b]REF refers to the reference genotype
[c]ALT refers to the alternate genotype

Supplemental Methods

Genome sequencing. Cells were grown on liquid media and genomic DNA extracted using Qiagen DNeasy Plant Maxi kits at OD_{680} = 0.8 followed by ethanol precipitation to concentrate DNA to 20 ng/uL. Paired-end libraries were constructed using Illumina TruSeq kits with 1 microgram of the concentrated genomic DNA. DNA was sheared using the Covaris sonicator with settings suggested in the TruSeq manual. Library preparation was carried out using agarose gel purification to select for insert sizes of approximately 400 bps. PCR amplification was performed for 10 cycles using Kapa HiFi DNA polymerase, which ensures higher fidelity in the GC-rich portions of *Chlamydomonas* DNA. Completed library quality was assessed on an Agilent Bioanalyzer using the DNA 1000 analysis kit, and concentration was determined by quantitative PCR (Kapa Biosystems Library Quantification Kit - Illumina/LightCycler® 480). 2 X 51 paired-end sequencing was performed on a HiSeq 2000 (Illumina). A small proportion of read pairs were discarded prior to alignment due to an imaging problem encountered for some samples.

Sequence alignment and processing. The *Chlamydomonas* reference genome (JGI v5) and annotation (v5.3.1) was downloaded from Phytozome (http://www.phytozome.net/) (Goodstein et al., 2011). The reference genome was modified to include chloroplast Genbank: BK000554.2 (Maul et al., 2002) and mitochondrial genome (Genbank: U03843.1) sequences. Raw HiSeq 2000 image data were processed with the Casava 1.8.2 (Illumina) pipeline and reads failing Illumina's default quality control filters

removed. Paired-end reads were aligned with the Burrows-Wheeler Aligner (BWA 0.6.1) (Li et al., 2008) aln and sampe programs with default settings. Sample BAM files were processed by running FixMateInformation (Picard-tools v. 1.62; http://picard.sourceforge.net), removing pairs of sequence reads where at least one read was marked as a duplicate by MarkDuplicates (Picard-tools), and re-aligning reads using the Genome Analysis Toolkit RealignerTargetCreator/IndelRealigner (GATK, version 2.6-4; DePristo et al., 2011) to minimize the number of SNPs called in indel regions. The sample alignments were then used in various downstream steps including TE insertion polymorphism and structural variant prediction. For SNP-calling, sample alignment files were merged with MergeSamFiles (Picard-tools v. 1.62) and sequence reads globally re-aligned across samples using the RealignerTargetCreator/IndelRealigner.

Variant calling and genotyping. SNP-calling was performed using the Unified Genotyper (UG) v.2.6-4 (DePristo et al., 2011) configured for haploid genomes. Reads with mapping quality zero and with low base quality were filtered prior to SNP calling per the GATK default settings. Base qualities were capped at their mapping quality, and bases close to indels adjusted during the SNP-calling step using the Base Alignment Quality (BAQ) method to reduce false positives near indels (Li and Durbin, 2009).

This yielded a set of 8,331,693 SNPs which were then filtered to reduce false positives including artefacts reported for deep sequencing data (Li, 2014). First, heterozygous genotypes in haploid samples are enriched for false positives (Li, 2014). To identify such problematic variants, SNPs were called with a parallel diploid run of the UG and variants with heterozygous genotype(s) excluded from further analysis. Second,

low complexity regions harbor a large proportion of artifactual SNPs and indels in high coverage data (Li, 2014). Approximately 8 Mb (7% of the genome assembly) of low complexity sequence was identified using mdust (http://bit.ly/mdust-LC) and SNPs in these regions were removed. Third, we partitioned the SNP data into bins based on individual summary statistics (e.g., depth, quality, degree of strand bias etc.) and estimated the ts:tv ratio in each bin. Tails of the distribution of each statistic were filtered if tails bins had unusually low ts:tv as these SNPs are likely enriched for false positives (DePristo et al., 2011; Liu et al., 2012). Finally, a small number of SNPs (~2,700) with non-reference base calls in the re-sequenced reference strain (CC-503) were filtered from our final call set. These SNPs represent either base-calling errors in the reference assembly or artifacts in the re-sequenced CC-503. Details of the filtering thresholds are listed in Supplemental Table 7.

Insertion/deletion (indel) polymorphisms were called using the UG configured for haploid genomes. Indels were filtered in a similar fashion to SNPs by excluding variants in low complexity regions (Li, 2014), excluding indels with heterozygote genotypes called in a diploid call set, and excluding indels with non-reference genotype calls in the re-sequenced CC-503. We also examined possible frameshift mutations, but the high number of frameshift calls, and their similar frequency in genes with *Arabidopsis* homologs versus those without suggested a high false positive rate for indel detection. We excluded indels from further consideration.

SNP effects were inferred with snpEff v. 3.5a (http://snpeff.sourceforge.net, Cingolani et al., 2012) including the -canon option. In cases where SNPs had multiple effect classifications, the effect predictions were simplified to the single most “damaging”

effect category (Cingolani et al. 2012) using the GATK VariantAnnotator (DePristo et al., 2011). A large percentage of *Chlamydomonas* codons segregate for multiple SNPs. The effect of such SNPs (e.g., nonsense, nonsynonymous or synonymous) can be mis-classified by snpEff which only considers the impact of individual SNPs in isolation. We therefore removed these SNPs from analysis of nonsense polymorphisms and summaries of SNP counts. However, to avoid excluding multi-SNP codons from diversity estimates, we adopted an evolutionary pathways approach (Nei and Gojobori, 1986) implemented using the software SNAP (Korber, 2000) to estimate nonsynonymous nucleotide diversity (π_N) and synonymous nucleotide diversity (π_S). Sites masked by our low complexity filter are not excluded in synonymous and nonsynonymous site counts in the π_N and π_S estimates. This is expected to have a minor impact on diversity estimates owing to the relatively small number of sites masked as low complexity in CDS regions (958,359/38,532,900 = 0.025).

Population genetic analysis. Population genetic parameters were estimated directly from the processed BAM alignments using POPBam v. 0.3 (Garrigan, 2013) in sliding windows. These window-based estimates of population parameters were therefore not subject to the filtering protocol described for SNP calls made by the Unified Genotyper. POPBam infers consensus bases at each genome position, applies a limited filtering protocol (e.g., eliminating calls from low coverage samples, requiring no missing data at a site), and is appropriate for population genetic inferences in haploid organisms. We inferred genome-wide levels of nucleotide diversity (Nei, 1987) and Kelly's ZnS (Kelly, 1997) statistics in sliding windows of 5, 10, or 25 kb. These summary statistics were

estimated by including one laboratory strain CC-125 (137c) as an additional independent sample, and excluding strain CC-2290 (a clonemate of CC-1952) (Gross et al., 1988) and CC-4414 (a field isolate which our analysis found to be almost identical to CC-125/CC-503). Inspection of polymorphism levels along chromosomes suggested that chromosome ends may have lower diverisity and higher LD compared with other genomic regions. We evaluated this by dividing the genome into intervals of 10 kb, estimating nucleotide diversity and Kelly's ZnS, and testing if the terminal-most 10 intervals (i.e., 100 kb) of each chromosome had different levels of diversity or LD than the remaining intervals with a two-tailed Wilcoxon Rank Sum Test. Principal component analysis (PCA) of SNP genotypes was conducted using SNPRelate (Zheng et al., 2012). The neighbor-joining (NJ) tree was constructed using a custom perl script to generate the Jukes-Cantor distance (Jukes and Cantor, 1969) matrix from the filtered SNP data and MEGA6 (Tamura et al., 2013) to implement the NJ method. The linkage disequilibrium statistic r^2 was inferred using the hap-r2 method in vcftools v 0.1.12a (Danecek et al., 2011).

STRUCTURE (v. 2.3.4, Pritchard et al. 2000) was used to cluster samples into populations. The input SNP dataset was prepared by randomly sampling approximately 10,000 SNPs from the filtered dataset described above to limit the impact of linkage on clustering (Pritchard et al. 2000). Analyses were conducted using the admixture model and Markov Chain Monte Carlo (MCMC) performed with a burn in of 150,000 steps and chain lengths of 350,000. Runs were conducted on 10 field isolates (excluding CC-4414) including only CC-1952 from the known clonemates (i.e., CC-2290/CC-1952) and one laboratory strain (CC-125) with correlated allele frequencies among populations, without

prior location information and assuming no linkage among SNPs. Analyses were repeated 10 times for each value of K (i.e., K=1 to K=6) to ensure predicted admixture proportions and model likelihoods were consistent among replicate runs. One run at K=3 yielded admixture proportions inconsistent with the other runs. This run and an arbitrarily selected run from each of the other K's were removed from further consideration (Supplemental Table 1). The Evanno method (Evanno et al. 2005) was implemented using a standalone version of Structure Harvester (version vA.1; Dent and vonHoldt 2012) to assist in identification of the optimal number of clusters. A spike in ΔK (Evanno et al. 2005) at K=3 (Supplemental Table 1) and the existence of largely unadmixed individuals representing each of the clusters at this value of K, but not K=4, suggest that three distinct populations exist in our sample set (Pritchard et al. 2000). Admixture proportions of individuals across multiple replicate runs were determined using CLUMPP (Jakobsson and Rosenberg 2007) and presented in Figure 2C.

Structural variation. Gene deletions were identified using the following coverage breadth criterion. Gene models with coverage breadth of at least 90% (i.e., 90% of sites covered by at least one read) in the resequenced reference strain, but less than 15% coverage breadth in at least one of the 19 non-reference strains were called deletions and those with less than 50% coverage were called partial deletions. For genes with multiple alternative splice variants, this criterion was applied to the canonical (i.e., longest) transcripts. Coverage breadth was determined using bedtools v. 2.17.0 (Quinlan and Hall, 2010). We were concerned that unusually high GC content (64%) in *Chlamydomonas* (Merchant et al., 2007) and 70% or higher in many coding regions might lead to low

coverage sequencing in gene regions and introduce false positives for gene deletions. This issue was addressed by requiring that gene models CC-503 be covered at 90% of nucleotide sites in order for the gene to be considered a deletion in a non-reference strain.

Regions with probable copy number gains were identified by calculating coverage per sample in 499 bp intervals using the GATK (DePristo et al., 2011) DepthOfCoverage tool. Normalized coverage was calculated as the $\log_2$ ratio of sample coverage per interval / median sample coverage across all windows. Large tracts (greater than approximately 30 kb) in which normalized coverage was approximately 2-fold higher than typical values were identified as duplications by manual inspection.

Additional classes of structural variation were predicted using a paired-end mapping (PEM) approach with SVDetect (version 0.8b, Zeitouni et al. 2010). For this analysis, sample alignments used for SNP-calling were further processed to remove read pairs where one or both reads had mapping quality less than 20 to reduce false positives associated with mis-mapping of reads (Lucas Lledó and Cáceres 2013). The processing script BAM_preprocessingPairs.pl (Zeitouni et al. 2010) was used to extract anomalous read pairs for input into the SVDetect linking program and to obtain σ length and μ length parameters from read pairs with normal forward/reverse orientations. SVDetect clusters were identified using window sizes calculated as $2*\mu + 2*\sigma$ and step length $(2*\mu + 2*\sigma)/4$ where μ and σ represent the mean and standard deviation of normally paired reads following the example in Zeitouni et al. (2010). Strand and order filtering were applied as described in Zeitouni et al. (2010) using μ and σ parameters as inputs and the minimum number of read pairs for a cluster to be retained was set to 20 read pairs. The σ threshold for each indel and duplication thresholds was set to 3. Predicted variants

associated with scaffolds, cpDNA, or mtDNA were removed from the final set of SV predictions. Additional filters were applied using default settings.

Transposable element (TE) insertion polymorphism. Transposable element insertions in sample genomes relative to the reference assembly were predicted using RetroSeq (Keane et al. 2013, https://github.com/tk2/RetroSeq). We used the soft-clipped, unaligned parts of the sequence reads covering the insertion sites to distinguish between independent insertions at closely situated sites. Retroseq identifies inconsistently mapped reads where one end is mapped with confidence (anchored reads), while it mate is either unmapped or mapped to a distant location with low mapping quality. In the discovery phase, discordant read pairs that may support a TE insertion were identified in each sample alignment (see above). We supplied a file specifying a set of TE types obtained from Repeatmasker (www.repeatmasker.org) and a corresponding BED file of locations in the reference genome using the option -refTEs. The output of the discovery phase is a list of read pair names per TE type were provided to the calling phase. The calling phase was run separately for each strain. For insertion predictions, we required a minimum mapping quality of 50 for the anchoring reads and only events supported by at least 24 read pairs including at least 12 forward oriented anchored reads upstream and 12 reverse oriented anchored reads downstream. We required that the distance from the last forward oriented upstream anchor to the first reverse oriented downstream anchor to be less than 120 bp. We used the –filter option and provided a set of TE predictions in the CC-503 reference TEs to limit redundancy of calls TE regions. The final call set was then filtered to include only the most confident TE insertion predictions with an FL tag value of 8 in the output VCF. Finally, we generate presence/absence matrix of the TE insertions in the

different *Chlamydomonas* strains (Supplemental Dataset 1) using bedtools (Quinlan and Hall, 2010)

Functional annotation. Gene Ontology (GO), Panther, PFAM, KEGG Orthology, and Arabidopsis homologs terms were retrieved from Phytozome version 9.0 (Goodstein et al., 2011) and KEGG Pathway assignments were obtained KEGG Orthology mapping (Kanehisa and Goto, 2000). Additional gene classifiers were obtained from specialized sources including GreenCut2 (Karpowicz et al., 2011), plantTFDB v3.0 (Zhang et al., 2011) (http://planttfdb.cbi.pku.edu.cn/), and Cildb v3.0 (Arnaiz et al., 2009) (http://cildb.cgm.cnrs-gif.fr/), which was the source for the flagellar proteome data of Pazour et al. (Pazour et al., 2005). GOSlim terms were obtained using the stand-alone version of the AgBase GOSlimViewer tool (http://www.agbase.msstate.edu/) (McCarthy et al., 2006). The *Chlamydomonas*-specific gene analysis and gene family-based analysis were based on predictions in the GreenPhylv4 database (http://www.greenphyl.org; (Rouard et al., 2011). Gene identifier mapping across genome versions were necessary to incorporate the GreenCut2 gene set into the analysis. Identifier mapping was performed using a table available from Phytozome 10 which is based on prediction from the Algal Functional Annotation Tool (Lopez et al., 2011). The v5 gene models in the Creinhardtii_236 reference annotation (Goodstein et al., 2011) used in our study had previously been filtered to remove genes overlapping > 30% overlap to transposable elements (Blaby et al., 2014). An additional 202 of 17,737 genes were found with TE-related PFAM domains (Piriyapongsa et al., 2007) and were excluded from the analysis.

All gene-level analyses are based on these 17,535 canonical transcripts (i.e., longest CDS) in the Phytozome version 9.0 annotation (i.e., Creinhardtii_236_gene.gff3).

***De novo* assembly and gene prediction**. Reads that failed to map to the reference genome were trimmed using Trimmomatic (Bolger et al., 2014) and *de novo* assembled using Velvet v. 1.2.08 (Zerbino and Birney, 2008). VelvetOptimiser (http://bioinformatics.net.au/software.velvetoptimiser.shtml) was used to identify the optimum k-mer lengths, expected coverage, and coverage cutoffs for each sample (VelvetOptimiser.pl -s 21 -e 45). Assembled contigs greater than 400 bps were retained and genes predicted with Augustus v. 2.7 (Stanke et al., 2004) incorporating *Chlamydomonas* species settings and allowing for partial gene models. Functional annotations were then predicted with InterProScan 5 (Jones et al., 2014) and a gene included in our *de novo* assembled gene set if the protein contained one or more PFAM domains found in eukaryotes (Phylodome, Novatchkova et al., 2005) and did not contain TE-related domains (Piriyapongsa et al., 2007). Predicted proteins were then confirmed by BLAST (Altschul et al., 1990) similarity searches against the non-redundant database (Pruitt et al., 2005) and manually scrutinized for potential contaminants. Proteins meeting these criteria were then searched against the CC-503 reference proteome using BLAST (Altschul et al., 1990) to verify they are unrelated to reference genes, and against the 345 kb mating type mt- locus (http://www.ncbi.nlm.nih.gov/nuccore/GU814015) (Ferris et al., 2010) to identify genes specific to the mt- locus that are absent from the CC-503 assembly. We note that the total assembled lengths of these contigs can exceed 10% of the CC-503 assembly.

SNP Validation. Thirty-three candidate SNPs were chosen from the set of predicted nonsense mutations for experimental validation. The target sequences were amplified by PCR, and the putative SNP checked by Sanger-based sequencing of one strain that has a reference allele genotype and at least one strain with the alternate allele. 31 of the 33 SNPs and the corresponding genotypes were confirmed to be correct using this approach (Supplemental Table 8).

Statistical analyses. GO-term enrichment analysis was performed using TopGO (Alexa and Rahnenfuhrer, 2010). Additional tests were performed using a two-tailed 2x2 Fisher's Exact Test or Chi-Square Test. All statistical analyses were performed using the R Statistical Programming Language v3.0 (http://www.r-project.org/).

Supplemental References

Alexa, A., and Rahnenfuhrer, J. (2010). topGO: Enrichment analysis for Gene Ontology. R package version 2.18.0.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol **215**, 403-410.

Arnaiz, O., Malinowska, A., Klotz, C., Sperling, L., Dadlez, M., Koll, F., and Cohen, J. (2009). Cildb: a knowledgebase for centrosomes and cilia. Database (Oxford) **2009**, bap022.

Blaby, I.K., Blaby-Haas, C.E., Tourasse, N., Hom, E.F., Lopez, D., Aksoy, M., Grossman, A., Umen, J., Dutcher, S., Porter, M., King, S., Witman, G.B., Stanke, M., Harris, E.H., Goodstein, D., Grimwood, J., Schmutz, J., Vallon, O., Merchant, S.S., and Prochnik, S. (2014). The Chlamydomonas genome project: a decade on. Trends Plant Sci **19**, 672-680.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**, 2114-2120.

Cingolani, P., Platts, A., Wang, l.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) **6**, 80-92.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., and Group, G.P.A. (2011). The variant call format and VCFtools. Bioinformatics **27**, 2156-2158.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., and Daly, M.J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet **43**, 491-498.

Earl, D.A., vonHoldt, B.M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetic Resources. **4**: 359-361.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol **14**, 2611-2620.

Garrigan, D. (2013). POPBAM: Tools for Evolutionary Analysis of Short Read Sequence Alignments. Evol Bioinform Online **9**, 343-353.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res **40**, D1178-1186.

Jakobsson, M., and Rosenberg, N.A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics **23**, 1801-1806.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.Y., Lopez, R., and Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics **30**, 1236-1240.

Jukes, T., and Cantor, C. (1969). Evolution of protein molecules. In: Mammalian Protein Metabolism. Munro, H. N. (eds). Illus. Academic Press: New York, N.Y., U.S.A. Vol. III. Xvii: 571p.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res **28**, 27-30.

Karpowicz, S.J., Prochnik, S.E., Grossman, A.R., and Merchant, S.S. (2011). The GreenCut2 resource, a phylogenomically derived inventory of proteins specific to the plant lineage. J Biol Chem **286**, 21427-21439.

Keane, T.M., Wong, K., and Adams, D.J. (2013). RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics **29**, 389-390.

Kelly, J.K. (1997). A test of neutrality based on interlocus associations. Genetics **146**, 1197-1206.

Korber, B. (2000). HIV signature and sequence variation analysis. In: Rodrigo AG, Learn GH. (eds) Computational Analysis of HIV Molecular Sequences. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics **30**, 2843-2851.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**, 1754-1760.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res **18**, 1851-1858.

Liu, Q., Guo, Y., Li, J., Long, J., Zhang, B., and Shyr, Y. (2012). Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics **13** Suppl 8, S8.

Lopez, D., Casero, D., Cokus, S.J., Merchant, S.S., and Pellegrini, M. (2011). Algal Functional Annotation Tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data. BMC Bioinformatics **12**, 282.

Lucas LledÛ, J.I., and C·ceres, M. (2013). On the power and the systematic biases of the detection of chromosomal inversions by paired-end genome sequencing. PLoS One **8**, e61292.

Maul, J.E., Lilly, J.W., Cui, L., dePamphilis, C.W., Miller, W., Harris, E.H., and Stern, D.B. (2002). The Chlamydomonas reinhardtii plastid chromosome: islands of genes in a sea of repeats. Plant Cell **14**, 2659-2679.

McCarthy, F.M., Wang, N., Magee, G.B., Nanduri, B., Lawrence, M.L., Camon, E.B., Barrell, D.G., Hill, D.P., Dolan, M.E., Williams, W.P., Luthe, D.S., Bridges, S.M., and Burgess, S.C. (2006). AgBase: a functional genomics resource for agriculture. BMC Genomics **7**, 229.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., MarÈchal-Drouard, L., Marshall, W.F., Qu, L.H., Nelson, D.R., Sanderfoot, A.A., Spalding, M.H., Kapitonov, V.V., Ren, Q., Ferris, P., Lindquist, E., Shapiro, H., Lucas, S.M., Grimwood, J., Schmutz, J., Cardol, P., Cerutti, H., Chanfreau, G., Chen, C.L., Cognat, V., Croft, M.T., Dent, R., Dutcher, S., Fern·ndez, E., Fukuzawa, H., Gonz·lez-Ballester, D., Gonz·lez-Halphen, D., Hallmann, A., Hanikenne, M., Hippler, M., Inwood, W., Jabbari, K., Kalanon, M., Kuras, R., Lefebvre, P.A., Lemaire, S.D., Lobanov, A.V., Lohr, M., Manuell, A., Meier, I., Mets, L., Mittag, M., Mittelmeier, T., Moroney, J.V., Moseley, J., Napoli, C., Nedelcu, A.M., Niyogi, K., Novoselov, S.V., Paulsen, I.T., Pazour, G., Purton, S., Ral, J.P., RiaÒo-PachÛn, D.M., Riekhof, W., Rymarquis, L., Schroda, M., Stern, D., Umen, J., Willows, R., Wilson, N., Zimmer, S.L., Allmer, J., Balk, J., Bisova, K., Chen, C.J., Elias, M., Gendler, K., Hauser, C., Lamb, M.R., Ledford, H., Long, J.C., Minagawa, J., Page, M.D., Pan, J., Pootakham, W., Roje, S., Rose, A., Stahlberg, E., Terauchi, A.M., Yang, P., Ball, S., Bowler, C., Dieckmann, C.L., Gladyshev, V.N., Green, P., Jorgensen, R., Mayfield, S., Mueller-Roeber, B., Rajamani, S., Sayre, R.T., Brokstein, P., Dubchak, I., Goodstein, D., Hornick, L., Huang, Y.W., Jhaveri, J., Luo, Y., MartÌnez, D., Ngau, W.C., Otillar, B., Poliakov, A., Porter, A., Szajkowski, L., Werner, G., Zhou, K., Grigoriev, I.V., Rokhsar, D.S., and Grossman, A.R. (2007). The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science **318**, 245-250.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol **3**, 418-426.

Nei, M. (1987). Molecular Evolutionary Genetics. New York: Columbia Univ. Press. pp 512.

Novatchkova, M., Wildpaner, M., Schweizer, D., and Eisenhaber, F. (2005). PhyloDome--visualization of taxonomic distributions of domains occurring in eukaryote protein sequence sets. Nucleic Acids Res **33**, W121-125.

Pazour, G.J., Agrin, N., Leszyk, J., and Witman, G.B. (2005). Proteomic analysis of a eukaryotic cilium. J Cell Biol **170**, 103-113.

Piriyapongsa, J., Rutledge, M.T., Patel, S., Borodovsky, M., and Jordan, I.K. (2007). Evaluating the protein coding potential of exonized transposable element sequences. Biol Direct **2**, 31.

Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics **155**, 945-959.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**, 841-842.

Rouard, M., Guignon, V., Aluome, C., Laporte, M.A., Droc, G., Walde, C., Zmasek, C.M., PÈrin, C., and Conte, M.G. (2011). GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Res **39**, D1095-1102.

Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res **32**, W309-312.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol **30**, 2725-2729.

Zeitouni, B., Boeva, V., Janoueix-Lerosey, I., Loeillet, S., Legoix-nÈ, P., Nicolas, A., Delattre, O., and Barillot, E. (2010). SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics **26**, 1895-1896.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18**, 821-829.

Zhang, H., Jin, J., Tang, L., Zhao, Y., Gu, X., Gao, G., and Luo, J. (2011). PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. Nucleic Acids Res **39**, D1114-1117.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics **28**, 3326-3328.