

1 **Common and rare genetic variation intersects with ancestry to** 2 **influence human skin and plasma carotenoid concentrations**

3 Yixing Han^{1*}, Savannah Mwesigwa^{2,6}, Qiang Wu⁴, Melissa N. Laska³, Stephanie B. Jilcott Pitts⁴,
4 Nancy E. Moran^{5*}, Neil A. Hanchard^{1,5,6*}

5 1) Center for Precision Health Research, National Human Genome Research Institute, National
6 Institutes of Health, Bethesda, MD

7 2) Department of Medical Microbiology, College of Health Sciences, Makerere University, Kampala,
8 Uganda

9 3) Division of Epidemiology & Community Health, School of Public Health, University of Minnesota,
10 Minneapolis, MN

11 4) Department of Public Health, East Carolina University, Greenville, NC

12 5) USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of
13 Medicine, Houston, TX

14 6) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX

15 *Corresponding Authors

16 Keywords: carotenoids, multi-ancestral, heritability, genetic variants, gene-by-dosage

17

18 **ABSTRACT**

19 Carotenoids are dietary bioactive compounds with health effects that are biomarkers of fruit and
20 vegetable intake. Here, we examine genetic associations with plasma and skin carotenoid
21 concentrations in two rigorously phenotyped human cohorts (n=317). Analysis of genome-wide
22 SNPs revealed heritability to vary by genetic ancestry ($h^2=0.08-0.44$) with ten SNPs at four loci
23 reaching genome-wide significance ($P<5E-08$) in multivariate models, including at *RAPGEF1*
24 (rs3765544, $P=8.86E-10$, $\beta=0.75$) with α -carotene, and near *IGSF11* (rs80316816, $P=6.25E-$
25 10, $\beta=0.74$), with cryptoxanthin; these were replicated in the second cohort (n=110). Multiple
26 SNPs near *IGSF11* demonstrated genotype-dependent dietary effects on plasma cryptoxanthin.
27 Deep sequencing of 35 candidate genes revealed associations between the *PKDIL2-BCO1* locus
28 and plasma β -carotene ($P_{adj}=0.04$, $\beta=-1.3$ to -0.3), and rare, ancestry-restricted, damaging
29 variants in *CETP* (rs2303790) and *APOA1* (rs756535387) in individuals with high skin
30 carotenoids. Our findings implicate novel loci in carotenoid disposition and indicate the
31 importance of including cohorts of diverse genetic ancestry.

32

33 INTRODUCTION

34 Carotenoids are a diverse group of natural pigments produced by plants, fungi, algae, and
35 photosynthetic bacteria. While there are over 1000 identified carotenoid species in nature¹, six
36 species (α -carotene, β -carotene, cryptoxanthin, lycopene, lutein, and zeaxanthin) constitute more
37 than 95% of total human blood carotenoids². Humans cannot synthesize carotenoids
38 endogenously, thus primarily acquire carotenoids from dietary fruits and vegetables (FV), mostly
39 in the form of β -carotene, lycopene, and lutein/zeaxanthin^{3,4}. Carotenoids are absorbed,
40 metabolized, and distributed throughout the blood, skin, and other tissues in a manner similar to
41 dietary lipids⁴. Carotenoid activities depend on their chemical properties, which can be pro-
42 vitamin A, nuclear receptor signaling, light filtering, or antioxidant/anti-inflammatory^{3,4}.
43 Because carotenoids are absorbed, retained in the body for a moderate amount of time, and are
44 detectable by spectroscopy, carotenoid concentrations, measured in plasma and more recently
45 non-invasively measured in the skin, have been proffered as biomarkers for dietary fruit and
46 vegetable intake assessment in adults and children^{5,6}.

47 Epidemiologic, clinical, and preclinical studies also indicate that carotenoids are
48 associated with protection from many chronic diseases, including cancers, cardiovascular
49 disease, and macular degeneration⁴. Their importance lies in their ability to modulate
50 intracellular signaling, offering antioxidant, antiapoptotic, and anti-inflammatory properties that
51 protect cells from oxidative stress, UV damage, and support functions like vision and immune
52 response³.

53 While carotenoid plasma and tissue concentrations are primarily a function of dietary
54 intake, there is still substantive inter-individual variation in plasma and tissue carotenoid
55 concentrations. Age, body mass index (BMI), and smoking have all been implicated as

56 environmental contributors to this inter-individual variation. Genetic variation is known to also
57 contribute to this variation^{4,7-10}. For instance, in Mexican American children, the heritability of
58 plasma carotenoid concentrations has been quantified, with α -carotene demonstrating a
59 heritability (h^2) of 0.81 ($P = 6.7 \times 10E-11$) and β -carotene exhibiting an even higher heritability
60 of 0.90 ($P = 3.5 \times 10E-15$)⁸. However, a detailed understanding of how genetic variation
61 influences interindividual remains unclear. A handful of early epidemiologic and clinical studies
62 found associations between common SNPs in select lipid and carotenoid metabolism genes and
63 blood concentrations of specific carotenoid species¹¹⁻¹⁵. An early genome-wide association study
64 (GWAS) in European populations identified an association with genetic variation near β -
65 *carotene 15,15-dioxygenase (BCO1)*¹², the key enzyme responsible for central cleavage of
66 provitamin A carotenoids to yield vitamin A¹⁶. At the tissue-specific level even less is known
67 about the impact of genetic variation on carotenoid levels; carotenoids in the macula of the eye
68 have been associated with SNPs in *BCO1*, *BCO2*, *NPC1L1*, *ABCG8*, and *FADS2*¹⁷⁻¹⁹, and in
69 small studies, carotenoids in prostate and skin have been associated with SNPs in the same
70 genes^{9,20}, albeit without subsequent replication of results.

71 To date, a broadly comprehensive understanding of how human genetic variation
72 influences plasma and tissue carotenoid concentrations, particularly bioactive carotenoid species,
73 remains elusive. This knowledge gap is strikingly evident for populations with non-European
74 ancestral backgrounds, as most prior studies have focused on individuals of European descent.
75 For example, early studies linking variants in *BCMO1* (aka *BCO1*) to plasma β -carotene levels¹²
76 and variants in *RBP4* (retinol-binding protein 4) to circulating retinol (a carotenoid metabolite)
77 levels²¹ were both conducted in homogenous Eurocentric cohorts. Additionally, *SCARB1* (a key
78 receptor for carotenoid uptake) has been associated with plasma lycopene levels in multiethnic

79 populations under certain conditions such as in postmenopausal women, though the effect sizes
80 vary across groups²². The lack of unbiased genome-wide analyses, particularly in ancestrally
81 diverse cohorts, restricts the generalizability of carotenoid genetics findings, limiting their
82 applicability in both population and precision nutrition strategies²³⁻²⁵. Conducting comprehensive
83 studies in diverse populations is thus essential going forward, as efforts to implement precision
84 medicine and nutrigenomic initiatives, and identify biomarkers that can be used across
85 population groups, intensify.

86 Here, we leverage detailed phenotypic data (including age, sex, BMI, self-reported
87 race/ethnicity, and food intake), and rigorous plasma and skin carotenoid assessments from two
88 ancestrally and geographically diverse US cohorts to conduct both genome-wide and targeted
89 (sequence-based) association studies of plasma carotenoid species and skin carotenoid levels. We
90 identify novel population common- and rare- genetic variants associated with steady-state and
91 diet-responsive carotenoid levels. Further, we reveal ancestry-specific differences in heritability
92 and genetic association. Using data derived from a controlled dietary intake study, we also
93 uncover gene-by-dosage interactions at associated loci. Collectively, our findings highlight
94 previously unrecognized genetic heterogeneity in human carotenoid metabolism, providing a
95 foundation for advancing precision nutrition and understanding global health and nutrigenetics.

96

97 **RESULTS**

98 Genetic studies were conducted in two extensively characterized, previously described
99 cohorts^{26,27} (**Methods**). The primary discovery cohort comprised 213 individuals (207 after QC
100 and familial relatedness check) from four self-reported United States (US) racial and ethnic
101 groups (Asian, Hispanic, Non-Hispanic Black, and Non-Hispanic White) (**Supplementary**

102 **Table S1**). Participants were recruited from two sites in the US, and had extensive clinical,
103 lifestyle, and demographic phenotype data documented alongside cross-sectional plasma
104 carotenoid species concentrations (measured by HPLC-photodiode array detection) and
105 aggregate skin carotenoids (skin carotenoid score - measured by non-invasive pressure mediated
106 reflection spectroscopy). These individuals were genotyped using the H3Africa genotyping
107 microarray (Illumina, CA, USA), designed to be used in genetically diverse populations. To
108 cover candidate loci that were not well-represented on the genotyping array, short-read capture-
109 based sequencing was performed across 35 genomic loci, which are reported to influence
110 carotenoid concentrations (**Methods, Supplementary Table S7**).

111 The secondary (intervention) cohort consisted of 162 individuals (**Supplementary Table**
112 **S2**), of whom 110 unique participants were not included in the primary cohort. Participants were
113 recruited from three sites, two of which were the same as the primary cohort, as part of a dietary
114 carotenoid intervention study. This study collected identical phenotypic data within a
115 longitudinal, randomized carotenoid-rich juice dose-response study. Data were collected at
116 baseline, 3-, and 6-weeks post-intervention for three mixed carotenoid doses (low/control (0
117 mg/d), moderate (4 mg/d), and high (8 mg/d)). This cohort underwent the same genomic
118 interrogation using the same platform and panel as the discovery cohort (**Figure 1**) (**Methods**
119 **and Supplementary Methods**). For genetic analyses, self-reported race/ethnicity was refined
120 through Multidimensional Scaling (MDS) of genomic information, aligning participants with
121 human ancestral populations based on the 1000 Genomes Project data (**Figure 2A**).

122

123 *Ancestral background influences carotenoid phenotype variability*

124 Carotenoid concentrations in plasma and skin are influenced by a variety of factors,
125 including indistinct biological factors proxied by self-reported racial-ethnic background^{27,28}.
126 Overall, we did not observe significant variation in the plasma carotenoid concentrations and
127 skin carotenoid levels among the four race/ethnicity groups in the Primary Study Cohort
128 (ANOVA); however, several significant differences in total plasma carotenoid levels were
129 observed among pairwise comparisons between the four primary cohort groups. Self-identified
130 non-Hispanic black and Asian individuals showed the most significant differences, including in
131 total plasma carotenoid (Welch's Two-Sample *t*-tests $P = 0.021$), plasma β -carotene ($P = 0.003$),
132 and plasma lutein/zeaxanthin ($P = 0.037$) (**Figure 2B; Supplementary Figure S2**). Additionally,
133 levels of α -carotene, lycopene, cryptoxanthin, and skin carotenoids differed significantly
134 between all four groups, with *t*-test *p*-values ranging from $8.67E-5$ (non-Hispanic black vs. Asian
135 for α -carotene) to 0.031 (non-Hispanic black vs. non-Hispanic white for lycopene) (**Figure 2** and
136 **Supplementary Figure S2**), while without significant differences between the EAS and SAS
137 (**Supplementary Figure S3**). Correlations between skin and plasma carotenoids were not
138 different between self-reported race and ethnicity²⁷; however, significant pair-wise differences in
139 skin carotenoid levels between self-reported groups mirrored observations in plasma carotenoids,
140 although the variance in the distributions of carotenoid species in groups was broader (**Figure**
141 **2C**).

142 The Primary Study Cohort included self-identified racial and ethnic groups consistent
143 with historical US census race and ethnicity categories; to appropriately contextualize these
144 groups for genetic studies, we aligned recruited individuals to genetic ancestry superpopulation
145 clusters from the 1000 Genomes Phase III dataset²⁹ using multi-dimensional scaling (MDS).
146 Most individuals identifying as 'white' and 'non-Hispanic black' clustered closely with 'European'

147 and 'African' genetic ancestry superpopulations, respectively. The reported 'Asian' race and
148 ethnic group, however, separated into two distinct clusters on the first two dimensions of the
149 MDS, with some individuals clustering with Indian/South Asian individuals (GIH) and others
150 aligning with East Asian ancestral groups (JPT, CHB) (**Figure 2A**). Individuals self-identifying
151 as 'Hispanic' clustered with mixed American and Hispanic ancestral groups (AMR). A small
152 group of individuals displayed notable discrepancies between their self-reported race/ethnicity
153 and genetic clustering (**Figure 2A**), including four individuals reported as 'white' and two
154 reported as 'Asian' whose genetic ancestry aligned more closely with 'South Asian' or 'Hispanic'
155 and 'AFR' super populations, respectively. (**Figure 2A**). For individuals discordant between self-
156 reported and genetically clustered ancestry, their genetic ancestry was used in downstream
157 analyses.

158

159 *The heritability of carotenoid concentration varies by carotenoid species and genetic ancestry*

160 To evaluate heritability in our cohort and facilitate downstream genetic analyses, we
161 derived a curated, quality-controlled (QC) dataset of 1,917,156 genotyped single nucleotide
162 polymorphisms (SNPs) from 207 healthy, unrelated individuals (**Supplementary Methods**).
163 This dataset was used to impute a total of 26,084,710 SNPs utilizing the Michigan Imputation
164 Server³⁰; with the 1000 Genomes Phase III v5 (GRCh37/hg19) reference panel serving as the
165 primary reference for this cohort of diverse US individuals. Further filtering for SNPs with a
166 correlation coefficient (r^2) of 0.3 or greater, and a minor allele frequency (MAF) greater than
167 0.05 resulted in a final dataset of 7,467,403 SNPs. Relative and absolute heritability of total and
168 sub-specified plasma carotenoids, as well as skin carotenoid levels, was then calculated in
169 GCTA³¹(**Methods**) using QCed autosomal SNPs.

170 In our primary cohort, the overall estimated heritability of plasma carotenoids was low
171 ($h^2=0.08$) albeit with a wide standard error ($se= 0.157$); perhaps unsurprising given the
172 substantial dietary contribution to the variance in carotenoid levels. Plasma lutein/zeaxanthin had
173 the lowest heritability estimate ($h^2=0.09$, $se= 0.172$) among carotenoid species, also reflecting a
174 heavy dietary influence. There was, however, notable variation in heritability estimates between
175 carotenoid species; for instance, plasma cryptoxanthin ($h^2=0.44$, $se= 0.322$) and α -carotene ($h^2=$
176 0.35 , $se= 0.338$) had higher heritability estimates compared to other species, which were
177 generally <0.17 (**Figure 3A, Supplementary Table S4**). By contrast, the heritability of skin
178 carotenoids was considerably higher than that of total plasma carotenoids ($h^2=0.08$ for plasma
179 carotenoids versus $h^2= 0.30$ for skin carotenoids) and more consistent with estimates for α -
180 carotene (**Figure 3A, Supplementary Table S4**).

181 To better understand the contribution of genetic ancestry to interindividual variation in
182 carotenoid phenotypes, we also estimated heritability across each genetic ancestry group. Given
183 the modest size of the resulting sub-samples, which resulted in large standard errors, we focused
184 on heritability estimates relative to the entire Primary Study Cohort. The heritability of total
185 plasma carotenoids was found to be relatively consistent across groups, except among
186 participants genetically clustering with Hispanic individuals (**Figure 3B**), in whom it was 9x
187 higher. A similar trend was observed for plasma α -carotene, β -carotene, and total lycopene,
188 which all exhibited higher relative heritability among Hispanic (AMR) clustering participants
189 (**Figure 3B**). By contrast, plasma cryptoxanthin had a relatively higher heritability (2.3x) among
190 African American clustering individuals (AFR), while the heritability of plasma lycopene was
191 notably higher among South Asian (SAS) clustering individuals (2.4x). The heritability of skin
192 carotenoids was highest among East Asian (EAS) clustering individuals (3.3x) (**Figure 3B**).

193 Generally, ancestry-specific relative heritability for skin carotenoids was higher among groups
194 outside of the European genetic cluster (**Figure 3B**).

195

196 *Common variants at novel loci are associated with plasma, but not skin, carotenoid*
197 *concentrations*

198 Given the relatively high heritability of some of the carotenoid species, we next sought to
199 identify genetic loci with significant effects on carotenoid concentrations across our diverse
200 cohort. Carotenoid measurements were log₂-transformed to provide a better approximation of a
201 normal distribution to be used in our statistical models (**Supplementary Table S3**). For most
202 measurements, such as plasma and food carotenoids, the log₂-transformed values showed
203 improved normality (e.g., for food carotenoids, $W = 0.794$; $P = 9.691E-16$ in the original data, W
204 $= 0.993$ and $P = 0.422$ after log₂ transformation) (**Supplementary Table S3**). However, for skin
205 carotenoid measurements, the original values had a better fit to normality. Using the transformed
206 plasma values, we first conducted a genome-wide association study (GWAS) of total carotenoids
207 and carotenoid species concentrations in plasma using linear regression models as implemented
208 in GEMMA and incorporating covariates of age, sex, BMI, log₂-transformed carotenoid intake,
209 and the first two MDS dimensions (**Supplementary Figure S1**).

210 In our analysis of plasma carotenoids, we identified six SNPs at three loci that reached
211 genome-wide significance ($P = 5E-08$) for either total- or plasma carotenoid species (**Table 1**,
212 **Figure 4**). A total of 37 SNPs at 12 loci surpassed a more permissive suggestive association
213 threshold ($P < 5E-06$). The strongest association was observed between plasma α -carotene
214 concentrations and rs3765544 (chr9:134458148:G>A) on chromosome 9q34 ($P = 8.86E-10$, beta
215 $= 0.750$; **Table 1, Figures 4B & 4G**), located in the intronic 24 region of the *RAPGEF1* gene.

216 This SNP had an effect size translating to an increase of 68% more α -carotene concentration per
217 allele. *RAPGEF1* encodes a guanine nucleotide exchange factor involved in the activation of *Ras*
218 family GTPases^{32,33} that plays a role in several cellular signaling pathways^{34,35,36}, and it is widely
219 expressed in tissues such as skeletal muscle and adipose tissue. Notably, the same SNP allele (G)
220 at this locus also showed marginal association ($P = 6.43E-08$, $\beta = 0.789$) with β -carotene
221 levels (**Table 1, Figure 4C & 4F Supplementary Table S5**), suggesting shared activity.

222 Consistent with the high heritability seen for plasma cryptoxanthin, we observed two
223 genome-wide significant loci associated with plasma cryptoxanthin concentrations: 1) multiple
224 SNPs downstream of *IGSF11* on chromosome 3q13 (**Figure 4A & 4D**; top SNPs:
225 chr3:118521532:T>C (rs80316816), $P=6.25E-10$, $\beta=0.08$; chr3:118494728:C>G
226 (rs76613159) $P=4.95E-09$, $\beta=0.08$; chr3:118446886:T>C (rs76087842) $P=8.32E-09$, $\beta =$
227 0.08) and 2) an intergenic locus on chromosome 19p13, comprising multiple SNPs (**Figure 4E**;
228 top SNPs: chr19:893793:C>A (rs28468554), $P=6.26E-09$, $\beta=-0.462$). Collectively, these two
229 loci account for approximately 32.95% of the variance in cryptoxanthin concentration. The 3q13
230 locus includes (within 50kb) genome-wide significant SNPs associated with education
231 attainment in two large studies^{37,38} and may not be regulatorily related to the closest gene
232 (*IGSF11*). In the Genotype-Tissue Expression (GTEx) database³⁹ the top SNP on 19p13
233 (rs28468554) is an expression quantitative trait locus (eQTL) SNP for *MEDI6* in multiple
234 tissues. *MEDI6* encodes a protein of the same name that is a component of the mediator
235 complex⁴⁰, which enables thyroid hormone and vitamin D3 receptor binding^{41,42}.

236 We used the same GWAS model to evaluate SNPs associated with skin carotenoids, but
237 for this analysis, we also included measurements of melanin and hemoglobin - both of which are
238 thought to influence skin carotenoid measurements²⁷ - as covariates. No variants surpassed either

239 the genome-wide or suggestive association threshold. We considered that by including skin tone
240 measurements and accounting for ancestry (as modeled in GEMMA), we may have
241 overcorrected for the ancestry effect; that is, if melanin and hemoglobin are collinear with some
242 genetic ancestries, also incorporating components reflecting genetic ancestry could be redundant.
243 Therefore, we reran the association using only clinical covariates (i.e. without MDS coordinates
244 or race/ethnicity) (**Supplementary Figure S4A**). This yielded 107 SNPs surpassing our genome-
245 wide significance threshold (**Supplementary Figure S4B**, genomic inflation factor = 2.06); this
246 suggested a strong effect of ancestry (population stratification) and was reflected in disparities in
247 minor allele frequencies at ‘associated’ loci between different genetic ancestry groups
248 (**Supplementary Figure S4D**). This disparity in association with and without ancestry
249 adjustments was not observed in the plasma carotenoid association analyses (**Supplementary**
250 **Figure S4C**).

251

252 *Replication and gene-by-dosage effects of carotenoid candidate SNPs*

253 To replicate our primary cohort findings, we conducted genome-wide genotyping in a
254 secondary cohort derived from a dietary carotenoid intervention study with a similar study
255 design²⁶ (**Methods**). This secondary cohort was smaller in size (n=162) and had a larger
256 proportion of Hispanic clustering (AMR, 23% vs. 14%), and a smaller proportion of European
257 clustering (EUR, 27% vs. 33%), individuals relative to our initial cohort (**Figure 1**). The
258 distribution of sex in the second cohort (male: n=79, 49%; female: n=83, 51%) is nearly equal,
259 contrasting with the predominantly female first cohort (male: n=62, 29%; female: n=151, 71%).
260 while age (median age = 29 years) was not different between the two cohorts (**Supplementary**
261 **Table S1 & S2**). Secondary cohort samples were genotyped on the same platform and underwent

262 identical quality control procedures as the primary cohort. We used baseline (pre-intervention)
263 plasma carotenoid and species measures as the outcome phenotype, applying the same GWAS
264 covariates and linear regression models in GEMMA with the 110 non-overlapping individuals
265 (i.e. only unrelated individuals who did not participate in the primary cohort were included in
266 analyses).

267 Meta-analysis between the two cohorts was then conducted in METAL⁴³ for the 37
268 suggestive threshold SNPs (representing 12 loci). We found nominal evidence for replication of
269 the same carotenoid species (cryptoxanthin) at the same two loci noted in our primary analysis
270 (*IGSF11* at 3q13 and *RNF111/MED16* at 19p13). Meta-analysis of rs1088589 at *IGSF11*
271 surpassed genome-wide significance, with a similar direction and magnitude of effect in both
272 cohorts ($p = 9.69E-08$, $\beta = 0.532$) (**Table 1**). Additionally, several other SNPs in this region (8
273 out of 12) surpassed the suggestive meta-analysis threshold ($P < 1E-06$) and exhibited the same
274 direction of effect. Two imputed SNPs, including our top SNP in *RAPGEF1*, were not observed
275 (imputed) in the secondary cohort and thus could not be replicated.

276 The interventional study design in our secondary cohort involved randomizing
277 participants to receive low, moderate, or high doses of dietary carotenoids via a daily,
278 carotenoid-enriched, fruit and vegetable juice, with skin and plasma carotenoid concentrations
279 measured at baseline (time 0) as well as 3- and 6-weeks post-intervention. This study design
280 allowed us to investigate whether SNPs at candidate loci identified in our primary cohort might
281 also influence the accumulation of carotenoids over time. We applied linear regression models
282 (**Methods**), incorporating the same covariates as in our initial study, along with additional terms
283 for intervention time points and a gene-by-dosage interaction effect (**Methods**). Of the 37
284 candidate SNPs identified in the discovery cohort and evaluated in the Intervention Cohort

285 **(Supplementary Table S5 and S6)**, 32 (86.5%) had significant F-statistic p-values (<0.05),
286 indicating a statistically significant relationship with the outcome variable (plasma carotenoid
287 concentration). Additionally, six (16.2%) SNPs – upstream of *IGSF11* (n=5) associated with
288 cryptoxanthin exhibited gene-by-dosage effects, where the effect of genotype on plasma
289 carotenoid concentrations differed by intervention dosage (**Figure 5A-F, Supplementary**
290 **Figure S5**).

291

292 *Target sequencing captures association between PKDIL2 and β -carotene concentrations*

293 Previous genome-wide studies have demonstrated an association between β -carotene and
294 common variants upstream of β -carotene oxygenase 1 (*BCOI*)¹². However, our primary GWAS
295 dataset lacked strong evidence for this. Whilst population differences (Eurocentric vs diverse
296 cohort) and sample size undoubtedly underlie part of this observation, we also noted low
297 coverage (few polymorphic SNPs) in this region on the genotyping array used; this is a known
298 limitation of array-based studies in genetically diverse populations^{44,45}. To address this, we
299 performed targeted sequencing of a subset of candidate genes (**Supplementary Table S7**)
300 identified from the literature^{4,46,47} and not adequately covered by the array. This sequencing
301 provided consistent median coverage across targeted loci (chr16:81101012-81220480), with
302 *PKDIL2*, upstream of *BCOI*, exhibiting a higher number of variants than other genes
303 (**Supplementary Figure S7**).

304 *PKDIL2* single nucleotide variants (SNVs) were the only SNVs from our targeted
305 sequencing cohort that were consistently associated with plasma or skin carotenoid levels in our
306 linear regression models, specifically with β -carotene concentrations. This association remained
307 significant when comparing the top third versus the bottom third of β -carotene concentrations

308 **(Supplementary Figure S7)**. *PKDIL2* variants also showed the strongest associations in the
309 replication cohort, with a similar direction (positive association) and effect sizes across the two
310 cohorts, though the specific variants differed **(Supplementary Figure S7)**. The *PKDIL2* locus is
311 found upstream of *BCO1*, in a region consistently associated with β -carotene concentrations.
312 Although our targeted capture did not directly sequence previously reported SNPs upstream of
313 *BCO1*¹¹, linkage disequilibrium (LD) patterns suggest that *PKDIL2* variants are likely to be in
314 the same LD block. As the *BCO1-PKDIL2* association was initially observed among the
315 European genetic ancestry group, we evaluated genotypes and β -carotene levels across our four
316 genetic ancestry groups. The top two SNPs (rs4148211 and rs7194871) were analyzed for
317 associations with β -carotene concentrations across four ancestry groups. A nominally significant
318 association was observed for rs4148211 in the Asian group (P=0.012, n=51), while no significant
319 associations were found for rs7194871 (p>0.5).

320

321 ***Rare, protein-damaging variants are observed in individuals with outlier carotenoid***
322 ***concentrations***

323 Rare, protein-damaging variants in coding regions of genes can have large effects on
324 physiologic traits⁴⁸. We, therefore, looked for rare, putatively protein-damaging variants
325 **(Supplementary Methods)** among individuals with extreme values (>2SD or <2SD) for either
326 plasma or skin carotenoids across Primary Study Cohort **(Supplementary Table S8)**.

327 In the Intervention Cohort, we identified an individual carrying a rare, predicted-
328 damaging, missense coding variant **(Supplementary Table S9)** (rs142824860,
329 NC_000016.9:g.81272554A>G; p.Glu14Gly; gnomAD MAF = 9.7E-05; CADD score 25.0)
330 **(Supplementary Methods)** in the *BCO1* gene who also had the lowest plasma β -carotene

331 concentrations in the cohort. For skin carotenoid concentrations, three notable outliers were
332 observed: two of these individuals clustered with the 1000 Genomes EAS population, and both
333 individuals carried a missense coding variant in *CETP* (rs2303790, NM_000078.3:c.1376A>G;
334 p.Asp459Gly; gnomAD MAF = 0.002) that is seen in 3% of East Asian (EAS) clustering
335 individuals but <1% in all other populations. The third outlier, clustering with other European
336 (EUR)-identifying individuals, possessed an ultra-rare (MAF=9.9E-06) variant in *APOAI*
337 (rs756535387; p.Arg201Ser; CADD score 24.0), predicted to be deleterious with an
338 AlphaMissense score of 0.701 (likely pathogenic) (**Supplementary Methods**). All three outliers
339 also had elevated plasma cryptoxanthin concentrations, which were strongly correlated with skin
340 carotenoid levels ($r = 0.57$; **Figure 6**).

341

342 **DISCUSSION**

343 Through GWAS and targeted sequencing analyses in two ancestrally diverse cohorts, we
344 provide comprehensive estimates of heritability and the genetic architecture underlying plasma
345 and skin carotenoid concentrations. We find that heritability varies across plasma carotenoid
346 species and is influenced by genetic ancestry, with estimates being generally lower for plasma
347 carotenoids compared to skin carotenoids. Notably, we identify and replicate an association
348 between genetic variation at chromosomes 3q13 (upstream of *IGSF11*) and 19p13 (*MED16*) and
349 plasma cryptoxanthin concentrations and find evidence of gene-by-dosage effects at the 3q13
350 locus. We confirm the association between β -carotene at the *PKDIL2-BCO1* locus and identify
351 putatively protein-damaging rare variants with large effects on skin and plasma carotene
352 concentrations.

353 We observed substantial heritability for plasma carotenoid species α -carotene and
354 cryptoxanthin, as well as skin carotenoids, with estimates comparable to those observed for
355 serum lipids such as cholesterol (ranging from 0.30 to 0.70)^{49,50}; this is consistent with the shared
356 biochemistry and metabolism of serum lipids and carotenoids. We also noted significant
357 variation in heritability estimates across different genetic ancestry groups, with individuals
358 clustering with admixed Amerindians (AMR; self-identified ‘Hispanic’) exhibiting high
359 heritability for nearly all plasma carotenoid sub-species, including replication of the high
360 heritability of α -carotene in previous studies among Mexican Americans ($h^2 = 0.85$ in this study;
361 0.81 in the previous study)⁸. Given that heritability represents the genetic contribution to total
362 variance, these results likely reflect the balance between genetic variation and differences in
363 typical dietary and environmental factors across groups. That some carotenoid species have
364 higher heritability estimates suggests that the metabolic events preceding their plasma
365 accumulation, such as intestinal absorption, are more sensitive to genetic variation. The public
366 health implication thereof is that dietary recommendations (or interventions) for certain
367 carotenoids (e.g. cryptoxanthin) aimed at achieving a specific plasma concentration range may
368 be more challenging to develop than for those carotenoid species with lower heritability. Larger
369 sample sizes will be important in further elucidating these differences and their public health
370 implications.

371 The distinct contributions of diet and genetics were further illustrated in our analyses of
372 skin carotenoid heritability. Non-invasive measurements of the tissue accumulation of
373 carotenoids in the skin present unique considerations. Principally, the efficiency of detecting
374 colorimetric changes due to carotenoid accumulation in the skin depends partially on skin
375 reflectivity, which may be influenced by skin melanin content and hemoglobin levels.

376 Previously, we demonstrated that, after adjusting for melanin, hemoglobin, and dietary intake, a
377 robust correlation remains between skin and plasma total carotenoid concentrations²⁷, though we
378 subsequently did not find skin melanin and hemoglobin to be significant modulators of skin
379 carotenoid responses to changes in dietary carotenoid intake¹⁰. Consequently, the apparent
380 heterogeneity in heritability estimates across ancestry groups could be a combination of
381 variability in confounding skin parameters and ancestral variability in factors influencing tissue
382 accumulation. The latter contention is worthy of consideration given the importance of
383 carotenoids to vitamin A metabolism, as carotenoids serve as precursors to retinoids, which are
384 critical for maintaining skin health⁴⁶. This highlights the potential influence of ancestral
385 variability on vitamin A-related physiological processes and the tissue accumulation of
386 carotenoids. Despite this, our understanding of the factors influencing tissue carotenoid
387 accumulation remains limited and warrants further investigation.

388 The relatively higher heritability for plasma cryptoxanthin was reflected in the number of
389 genome-wide- and suggestive associations compared to concentrations of other carotenoid
390 species, especially at the 3q13 and 19p13 loci. Few if any previous genetic studies have
391 considered cryptoxanthin concentrations, and to the best of our knowledge, a role for genetic
392 variation at these two loci has not previously been described. At 3q13, the closest gene, *IGSF11*,
393 encodes a member of an immunoglobulin superfamily⁵¹ whose primary function is as a cell
394 adhesion molecule that stimulates cell growth^{52,53}; however, in addition to neurocognition,
395 genetic variation at 3q13 has been implicated in gut microbiota diversity^{54,55} and body mass
396 index⁵⁶⁻⁵⁸, suggesting that the full functional regulatory impact of this locus may not yet be well
397 understood. The 3q13 locus also provided five of the six SNPs (all in strong linkage
398 disequilibrium (LD) with each other) with evidence for gene-by-dosage effects, with the mutant

399 (non-reference) minor allele being negatively correlated with cryptoxanthin concentrations at
400 high FV doses but being positively correlated or neutral at low or intermediate FV doses. We
401 documented the gene-by-dosage effect for dietary carotenoid intervention and further highlighted
402 the complexity of making personalized nutrition recommendations for specific carotenoid
403 species. Cryptoxanthin, while a relatively small component of total plasma carotenoids, serves as
404 a highly bioavailable vitamin A precursor, and confers reduced inflammation, improves immune
405 function, and antioxidant activity⁵⁹⁻⁶¹. A recent longitudinal population study found a strong
406 positive association between maternal cryptoxanthin concentrations at delivery and offspring
407 cognitive development at age two⁶².

408 Targeted sequencing further enhanced our ability to capture the full spectrum of genomic
409 variation, particularly given the diverse ancestries included in our study. This approach
410 facilitated the interrogation of loci that are not well captured or imputed in diverse cohorts using
411 genotyping microarrays, and the identification of rarer and novel variants that would not be
412 detectable from fixed-content arrays. This was most evident at the *PKDIL2-BCO1* locus; these
413 two genes are arranged in reverse tandem (head to tail) within an ~18 kb stretch on chromosome
414 16q. Common variants in this region – ranging from the 5'upstream of *BCO1* to *CETP* have been
415 consistently associated with β -carotene metabolism^{11,12,63}; however, narrowing down putatively
416 causal variants in this region has been elusive. Our findings underscore this uncertainty –the top
417 associated variants were in the fifth exon of *PKDIL2*, but the top associated variant was much
418 further downstream of *PKDIL2* in our Intervention Cohort (**Supplementary Figure S7**). The
419 challenge of replicating individual variants at this locus across studies likely stems from
420 differences in population ancestry, environmental factors (such as adequately accounting for
421 dietary carotenoid intake), and study design, all of which may distort the association of variants,

422 particularly if multiple associated alleles each have small effect sizes. Regardless, the gene-level
423 association is sufficiently consistent that the *PKDIL2* may harbor multiple common variants,
424 each contributing modestly to β -carotene levels, collectively exerting a significant effect. In line
425 with this, we identified a rare, damaging variant (rs142824860) in the *BCO1* gene in our
426 Intervention Cohort in an individual with the lowest β -carotene levels, further highlighting the
427 potential for multiple variants with varying effects to contribute to population levels of β -
428 carotene¹⁰.

429 Our results also suggest a strong putative overlap between lipid metabolism and
430 physiological carotenoid regulation. Six genes (*ALDH7A1*, *ATF6*, *MED16*, *SALL1*, *SORBS1*,
431 *SORBS2*) near our plasma carotenoid suggestive candidate loci, as well as both genes (*CETP*
432 and *APOA1*) harboring high-impact rare variants in individuals with outlier carotenoid
433 concentrations, are either known or suspected modulators of lipid metabolism⁶⁴⁻⁶⁹. Among these,
434 the *ATF6* locus on chromosome 1 had the strongest statistical association, with the rs11579627
435 SNP (chr1:161930954:G>A) nearing genome-wide significance (P = 8.23E-08, beta = 0.32),
436 **Supplementary Table S5**). Notably, *ATF6*, a key transcription factor in the endoplasmic
437 reticulum (ER) stress response and the unfolded protein response (UPR) pathway⁷⁰, is implicated
438 in lipid biosynthetics⁶⁵. These associations highlight allelic variation in genes predominantly
439 involved in cell signaling and lipid metabolism.

440 The missense coding variant in *CETP* has been previously linked to exudative age-related
441 macular degeneration^{71,72} (a condition related to carotenoid nutrition) and elevated high-density
442 lipoprotein cholesterol levels (a determinant of plasma carotenoid concentrations)⁷³. Whilst the
443 overlap between lipid and carotenoid loci is not entirely surprising, given similarities in the
444 biochemistry of both, it does suggest that identifying genetic contributors to carotenoid

445 concentrations in larger studies could benefit from overlapping a comprehensive compendium of
446 lipid metabolism genes and that genetic studies of lipid variation, particularly in diverse
447 populations, would benefit from including carotenoid assessments and using colocalization to
448 identify strong biological candidates.

449 The modest sample sizes and relatively balanced distribution of ancestry groups in our
450 study means that our study was necessarily aimed at identifying loci with large trans-ancestry
451 effects that are likely to be relevant across ancestries. Conducting our analysis in a diverse multi-
452 ancestry cohort, however, still provided unique insights that would not have been evident using a
453 more ancestrally homogenous study group, particularly as it pertains to the allelic spectrum
454 underlying carotenoid variability. For instance, most of the carotenoid heritability estimates were
455 relatively higher among non-European populations, and all of the suggestive and genome-wide
456 significant candidate variants observed had higher minor allele frequencies in non-European
457 populations; this was particularly true for skin carotenoid concentrations. The inclusion of
458 diverse genetic ancestry in our rare variant studies further emphasized the utility of ancestrally
459 diverse cohorts – the coding missense variant in *CETP* (rs2303790) associated with very high
460 skin and cryptoxanthin levels is predominantly common among individuals with East Asian
461 ancestry.

462 Despite the insights gained from our analysis, there are limitations to our study. Our
463 sample size is small in comparison to modern GWAS; whilst this undoubtedly limited our power
464 to detect variants/loci with smaller effect sizes, our sample size is comparable to that used to
465 discover major effect loci for more commonly measured physiologic proteins (e.g. fetal
466 hemoglobin levels and cholesterol), and replicating our results in a second independent cohort
467 mean that the reported associations are unlikely to be false positives. Despite the well-

468 documented health effects of carotenoids, carotenoid concentrations are not routinely measured
469 clinically or included in large-scale biobanks and databases; as a result, these resources are not
470 available to further replicate our findings. Additionally, although supplemented by imputation
471 and, for some loci, targeted sequencing, our reliance on genotyping arrays, particularly in a
472 cohort with diverse ancestries, may have missed rare or novel variants that could either be
473 independently associated with carotenoid concentrations or augment findings at suggestive loci.
474 Potential differences in population structure (e.g. linkage disequilibrium) and/or nutritional
475 factors between the discovery and replication cohorts may have contributed to a lack of
476 replication of some discovery associations, especially if multiple associated alleles each have
477 small effect sizes.

478 Going forward, there are several lessons for future carotenoid and nutrigenetic research.
479 Principally, from a genetic standpoint, larger and more diverse cohort studies of carotenoids are
480 necessary to replicate our findings and enhance the robustness and generalizability of the
481 identified associations. Methods of incorporating local genetic ancestry⁷⁴ and deconvoluting
482 ancestry effects in trans-ancestry GWAS continue to improve and are likely to be particularly
483 important for exploring and refining carotenoid associations, given the variability noted across
484 ancestry groups. Additionally, incorporating detailed phenotyping of potential covariates,
485 controlled interventions, and measurements of lipids and related physiological compounds is
486 likely to be fruitful in understanding the complex underlying physiology. The rare variant
487 candidates identified here provide strong starting points for *in vitro* and *in vivo* functional
488 validation studies, particularly at the *CETP* (rs2303790) and *BCOI* genes. Finally, large
489 population-based studies would provide the necessary data to consider developing personalized
490 risk profiles that incorporate carotenoid (and related compound) measurements, genetic factors,

491 and independent demographic and dietary interactions. Such studies have the potential to provide
492 the level of detail needed to tailor public health recommendations for FV intake and
493 interventions across different population groups.

494 The comprehensive, agnostic view of the genetics of carotenoid metabolism presented
495 here provides a robust starting point for future studies of this important class of natural dietary
496 compounds and underscores the necessity of including diverse ancestry groups and deep
497 phenotyping in precision nutrigenetic research going forward.

498

499 **METHODS**

500 **Study design and carotenoid species measurement.**

501 We utilized the samples and phenotypic data collected from individuals who participated in two
502 previous studies^{26,27}. In the first study, participants were healthy adults aged 18-65 years,
503 recruited from two sites in North Carolina and Minnesota. They self-identified as Non-Hispanic
504 Black or African American (hereafter referred to as Non-Hispanic Black), Asian, Non-Hispanic
505 White, or Hispanic. The demographics of the Primary Study Cohort are detailed in

506 **Supplementary Table S1**. Skin carotenoids were measured using pressure-mediated reflection
507 spectroscopy (Veggie Meter, Longevity Link, Utah), which returns and aggregates skin
508 carotenoid score measurements that correspond with multiple skin carotenoids^{75,76}. Total plasma
509 carotenoid concentrations were determined via an HPLC-photodiode array²⁶. The resulting
510 dataset comprised SNPs from a cohort with the following self-identified racial and ethnic
511 distribution: non-Hispanic black (61, 29%), Asian (53, 25%), non-Hispanic white (70, 33%), or
512 Hispanic (29, 14%). Participants were predominantly female (N=151 (71%)), with fewer males
513 (N=62 (29%)), and a median age of 30 years.

514 The second cohort was also drawn from a previous intervention study²⁶, recruited from three
515 sites in North Carolina, Minnesota, and Texas. The racial and ethnic distribution of the
516 Intervention Cohort slightly differed from the primary cohort, consisting of non-Hispanic black
517 (41, 25%), Asian (40, 25%), non-Hispanic White (44, 27%), and Hispanic (37, 23%)
518 participants, with a near-equal sex distribution (49% male and 51% female) (**Supplementary**
519 **Table S2**). Plasma and skin carotenoid concentrations were collected at three time points
520 (baseline/week 0, week 3, and week 6) using the same methods. Participants were randomized to
521 receive negligible, medium-dose (4 mg total carotenoids/day), or high-dose (8 mg total
522 carotenoids/day) of dietary carotenoids for the intervention. Daily intervention adherence was
523 recorded by participants and non-intervention carotenoid intake was assessed with repeated 24-
524 hour dietary recalls prior to each visit.

525 **Multidimensional Scaling (MDS)**

526 Next, we combined our dataset with the 1000 Genomes phase III data to estimate the genetic
527 ancestry of the cohort. The 1000 Genomes phase III data were acquired from
528 <https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/> following the instructions at
529 https://www.cog-genomics.org/plink/2.0/resources#1kg_phase3. After downloading, the data
530 was converted to PLINK binary format, removing ambiguous SNPs (i.e. A>T/T>A and
531 C>G/G>C SNPs that are indistinguishable at the strand level), non-AT, and non-GC SNPs. The
532 data was pruned to remove SNPs with $R^2 > 0.1$ in windows of 50 SNPs advancing 10 SNPs at a
533 time across the chromosome (--indep-pairwise 50 10 0.1), after which SNP nucleotide
534 mismatches were corrected and merged with our cohort data using similar QC filters.
535 Subsequently, Multidimensional Scaling (MDS) was performed using the --cluster and --mds-
536 plot options in PLINK, generating eigenvalues and eigenvectors that encapsulate the MDS

537 dimensions and their respective scores for each individual. R version 4.3 was then used for
538 plotting the MDS and scree plot.

539 **Genome-wide heritability analysis**

540 Heritability estimates were calculated using Genome-wide Complex Trait Analysis (GCTA)³¹
541 version 1.94.1. The input data for the analysis comprised quality-controlled, genotyped, genome-
542 wide SNPs with minor allele frequency (MAF) thresholds of 0.01(--autosome --maf 0.01). Total
543 plasma carotenoids, plasma species, and skin carotenoid measurements for each individual
544 included in the filtered genotyping dataset served as the phenotypic data for the heritability
545 assessment. Genetic ancestry was used for downstream genomic analyses. Relative heritability is
546 calculated as the heritability estimate in each subgroup divided by the heritability estimate in the
547 entire cohort.

548 **Genome-Wide Association Studies (GWAS)**

549 Genome-wide association studies (GWAS) were conducted to explore the relationship between
550 imputed genotyping variants and carotenoid species levels, employing GEMMA⁷⁷ version
551 0.98.5. As an orthogonal assessment, we also conducted association using PLINK 1.9; all
552 primary results reported relate to tests done in GEMMA. For analyses of plasma carotenoid and
553 species covariates of age, sex, BMI, race/ethnicity (self-reported or MDS-defined), and food
554 carotenoid were used. Additionally, for the skin carotenoids association study, the melanin index
555 and hemoglobin index were included, consistent with previous findings²⁸.

556 Before association testing, the normality of raw data, log₂-transformed phenotypes, and
557 covariates was assessed using the Shapiro-Wilk Test (**Supplementary Table S3**). Measurements
558 approximating a Gaussian distribution were included in the association analysis. Log₂-
559 transformed values generally improved normality for plasma and food carotenoids, while

560 original values better fit skin carotenoid data. GWAS was then conducted on both the study and
561 intervention cohorts.

562 **Replication and transferability of significant SNPs**

563 Suggestively associated variants ($P = 5E-04$) with plasma carotenoid species and skin
564 carotenoids, identified from the GWAS in the Primary Study Cohort, were further evaluated in
565 the Intervention Cohort. We conducted an association analysis on individuals who participated
566 exclusively in the Intervention Cohort ($n=110$) using GEMMA, applying models similar to those
567 used in the Primary Study Cohort. Baseline carotenoid measurements served as the phenotypic
568 data. METALysis⁴³ was used with significantly associated SNPs in this Intervention Cohort.
569 For the second replication analysis, changes in plasma carotenoid species and skin carotenoid
570 levels from baseline (week 0) to the intervention endpoint (week 6) were used as phenotypes,
571 following the approach of a previous study²⁷. Covariates including age, sex, BMI, baseline
572 carotenoid concentrations, treatment assignment (0, 1, 2), study sites, and the first and second
573 dimensions of MDS together with SNPs and SNPs * treatment assignment (0, 1, 2) were fitted
574 into a linear regression model. For skin carotenoids, we also incorporated the melanin and
575 hemoglobin index in the linear regression model.

576 **Linear regression model**

- 577 • Y = Change in plasma carotenoid species or skin carotenoid levels
- 578 • β_0 = Intercept
- 579 • $\beta_1, \beta_2, \beta_3, \dots, \beta_n$ = Coefficients for each covariate
- 580 • X_1 = Age
- 581 • X_2 = Sex (coded as a binary variable)
- 582 • X_3 = BMI
- 583 • X_4 = Baseline carotenoid levels

- 584 • X_5 = Treatment assignment (0, 1, 2)
- 585 • X_6 = Study site (coded as necessary)
- 586 • X_7 = First MDS coordinate
- 587 • X_8 = Second MDS coordinate
- 588 • X_9 = Melanin index (for skin carotenoids)
- 589 • X_{10} = Hemoglobin index (for skin carotenoids)
- 590 • SNP_i = Genotypic data for the i-th SNP (coded as necessary)
- 591 • $SNP_i \times Treatment$ = Interaction term between SNP and treatment assignment

592 The linear regression model can be expressed as:

$$593 \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} \\ 594 \quad + i \sum \beta_{SNP_i} SNP_i + j \sum \beta_{SNP_j \times Treatment} (SNP_j \times Treatment) + \epsilon$$

595 Where:

- 596 • ϵ = Error term representing the variability not explained by the model.

597

598 **Targeted sequencing and bioinformatics analysis**

599 Thirty-five (35) genes (**Supplementary Table S7**) identified as important for carotenoid
600 metabolism were sequenced using the Ampliseq Custom Panel from Illumina. FASTQ files
601 were aligned to the GRCH38/hg38 reference genome using BWA⁷⁸ MEM with the following
602 parameters: -M -O 30 -E 4 -T 20 -v 3. Data from multiple lanes were then merged and indexed
603 using SAMtools⁷⁹ v1.5. Per-base coverage was computed with bedtools⁸⁰ coverage
604 (**Supplementary Table S7, Supplementary Figure S6**). Subsequently, the Genome Analysis
605 Toolkit (GATK)⁸¹⁻⁸³ v4 was utilized for variants calling following the steps of
606 *MarkDuplicatesSpark*, *BaseRecalibrator*, *ApplyBQSR*, *HaplotypeCaller*, *GenomicsDBImport*
607 and *GenotypeGVCFs*. SNP annotation was performed using ANNOVAR⁸⁴ (version 2020-06-08).

608

609 **Targeted sequencing association study**

610 The Variants Association Tool⁸⁵ and VCFTools⁸⁶ (v0.1.15) were employed to filter the variants
611 within the Ampliseq target regions. Quality control filters were applied to exclude variants with a
612 minor allele frequency (MAF) below 0.01 and a call rate below 95%. Variants significantly
613 deviating from Hardy-Weinberg equilibrium ($p < 0.001$) were also excluded. We focused on
614 non-synonymous variants by filtering out synonymous substitutions, thereby retaining only those
615 variants with potential impacts on protein function for downstream analysis.

616

617 **Gene-by-dosage effect analysis**

618 To evaluate the interaction between genetic variants and intervention dosage on plasma
619 carotenoid levels, we performed a Gene-by-dosage analysis using the linear regression model
620 above. For each locus, genotype-by-treatment interactions were assessed by fitting a model
621 where the \log_2 -transformed plasma carotenoid levels at the intervention endpoint (week 6) were
622 the dependent variable. The independent variables included the genotype of the specific locus,
623 intervention dosage group (control, moderate, or high), baseline plasma carotenoid levels, and a
624 range of covariates: age, sex, BMI, study site, and race. We also included the interaction term
625 between genotype and treatment dosage. This model was applied across loci, allowing for locus-
626 specific examination of the gene-by-dosage effect. The significance of the gene-by-dosage effect
627 was determined by the p-value, with a significance threshold of $p < 0.1$. Results were visualized
628 using boxplots with fitted regression lines for each dosage group, illustrating the differential
629 effects of intervention dosage across genotypes.

630

631 **DATA AVAILABILITY**

632 H3Africa Array - <https://h3africa.org/index.php/2019/12/12/h3africa-chip-faq/>

633 H3Africa Array Data Sheet - [https://www.illumina.com/downloads/infinium-h3africa-](https://www.illumina.com/downloads/infinium-h3africa-consortium-array-data-sheet-370-2020-001.html)
634 [consortium-array-data-sheet-370-2020-001.html](https://www.illumina.com/downloads/infinium-h3africa-consortium-array-data-sheet-370-2020-001.html)

635 1000 Genomes Project Resource - <https://www.internationalgenome.org/>

636 The datasets generated during and/or analyzed during the current study are available from the
637 corresponding author upon reasonable request.

638

639 **ACKNOWLEDGEMENTS**

640 The authors would like to thank all the persons who participated in data acquisition and sharing
641 during the two cohorts' establishment. This research was supported by the NIH NHLBI (SJP,
642 MNL, NEM, QW: 1R01HL142544-01A1), by funding from NIH NHGRI (HG-200412 to
643 N.A.H.), and the USDA/ARS (cooperative agreement 3092-51000-059-002S to NEM). The
644 work and views expressed do not reflect the views of the NIH or the USDA. This work utilized
645 the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>).

646

647 **AUTHOR CONTRIBUTIONS**

648 N.E.M. and N.A.H. designed the study. M.N.L., N.E.M., and S.B.J.P. led the teams that recruited
649 participants, obtained informed consent, and collected samples. S.M. performed the rare variant
650 association analysis. Y.H. conducted all other data analyses, including ancestry assessment,
651 heritability estimation, GWAS, statistical genetics, and bioinformatics processing of the target
652 sequencing data. Q.W. provided support for statistical analysis. Y.H. wrote the initial first draft,

653 with intellectual content added by S.M., N.A.H., and N.E.M. All authors reviewed and approved
654 the final manuscript.

655

656 **DECLARATION OF INTERESTS**

657 The authors do not have any conflicts or relevant interests to declare.

658

659 **SUPPLEMENTARY INFORMATION**

660 **Supplementary Tables**

661 Supplementary Table S1. Demographics of the Primary Study Cohort.

662 Supplementary Table S2. Demographics of the Intervention Cohort.

663 Supplementary Table S3. Normality test of the phenotypic measurements.

664 Supplementary Table S4. Estimated heritability for carotenoid species.

665 Supplementary Table S5. SNPs that are significantly associated with carotenoid metabolism.

666 Supplementary Table S6. Results of the linear regression test for gene-by-dosage interactions
667 across 37 SNPs.

668 Supplementary Table S7. Per-base coverage of 35 genes from target sequencing.

669 Supplementary Table S8. *PKDIL2* variants associated with plasma β -carotene.

670 Supplementary Table S9. Variants associated with plasma cryptoxanthin, skin carotenoid levels,
671 and β -carotene identified in outlier analysis.

672 **Supplementary Figures**

673 Supplementary Figure S1. Scree plot of the MDS eigenvalues for the Primary Study Cohort.

674 Supplementary Figure S2. Density plots of plasma carotenoid subspecies concentrations.

675 Supplementary Figure S3. Density plots of plasma carotenoid and subspecies concentrations (and
676 skin carotenoids in the East Asian (EAS) and South Asian (SAS) groups.

677 Supplementary Figure S4. Genome-wide association analysis of skin carotenoid level.

678 Supplementary Figure S5. Gene-by-dosage plots of plasma carotenoids, genotype, and
679 intervention dosage at week 6.

680 Supplementary Figure S6. Distribution and effects of genetic variants across selected genes.

681 Supplementary Figure S7. LocusZoom plot of variants in the *PKDIL2* gene on chromosome 16.

682 **Supplementary Methods**

683 DNA processing and genotyping

684 Genotyping data imputation

685 SNP and individual quality control

686 Rare Variants Annotation

687 **REFERENCES**

- 688 1. Yabuzaki, J. Carotenoids Database: structures, chemical fingerprints and distribution among
689 organisms. *Database (Oxford)* **2017**(2017).
- 690 2. Maiani, G. *et al.* Carotenoids: actual knowledge on food sources, intakes, stability and
691 bioavailability and their protective role in humans. *Mol Nutr Food Res* **53 Suppl 2**, S194-218
692 (2009).
- 693 3. Kaulmann, A. & Bohn, T. Carotenoids, inflammation, and oxidative stress--implications of
694 cellular signaling pathways and relation to chronic disease prevention. *Nutr Res* **34**, 907-29
695 (2014).
- 696 4. Moran, N.E., Mohn, E.S., Hason, N., Erdman, J.W., Jr. & Johnson, E.J. Intrinsic and Extrinsic
697 Factors Impacting Absorption, Metabolism, and Health Effects of Dietary Carotenoids. *Adv Nutr*
698 **9**, 465-492 (2018).
- 699 5. Ermakov, I.V. *et al.* Skin Carotenoids as Biomarker for Vegetable and Fruit Intake: Validation of
700 the Reflection-Spectroscopy Based "Veggie Meter". *Faseb Journal* **30**(2016).
- 701 6. Campbell, D.R. *et al.* Plasma carotenoids as biomarkers of vegetable and fruit intake. *Cancer*
702 *Epidemiol Biomarkers Prev* **3**, 493-500 (1994).
- 703 7. Tremblay, B.L., Guenard, F., Lamarche, B., Perusse, L. & Vohl, M.C. Genetic and Common
704 Environmental Contributions to Familial Resemblances in Plasma Carotenoid Concentrations in
705 Healthy Families. *Nutrients* **10**(2018).
- 706 8. Farook, V.S. *et al.* Genetics of serum carotenoid concentrations and their correlation with
707 obesity-related traits in Mexican American children. *Am J Clin Nutr* **106**, 52-58 (2017).
- 708 9. Norman, A.C., Palmer, D.G., Moran, N.E., Roemmich, J.N. & Casperson, S.L. Association of
709 Candidate Single-Nucleotide Polymorphism Genotypes With Plasma and Skin Carotenoid
710 Concentrations in Adults Provided a Lycopene-Rich Juice. *J Nutr* **154**, 1985-1993 (2024).
- 711 10. Bohn, T. *et al.* Host-related factors explaining interindividual variability of carotenoid
712 bioavailability and tissue concentrations in humans. *Mol Nutr Food Res* **61**(2017).
- 713 11. Hendrickson, S.J. *et al.* beta-Carotene 15,15'-monooxygenase 1 single nucleotide polymorphisms
714 in relation to plasma carotenoid and retinol concentrations in women of European descent. *Am J*
715 *Clin Nutr* **96**, 1379-89 (2012).
- 716 12. Ferrucci, L. *et al.* Common variation in the beta-carotene 15,15'-monooxygenase 1 gene affects
717 circulating levels of carotenoids: a genome-wide association study. *Am J Hum Genet* **84**, 123-33
718 (2009).
- 719 13. Borel, P., Desmarchelier, C., Nowicki, M. & Bott, R. Lycopene bioavailability is associated with
720 a combination of genetic variants. *Free Radic Biol Med* **83**, 238-44 (2015).
- 721 14. Borel, P. *et al.* Interindividual variability of lutein bioavailability in healthy men:
722 characterization, genetic variants involved, and relation with fasting plasma lutein concentration.
723 *Am J Clin Nutr* **100**, 168-75 (2014).
- 724 15. Borel, P., Desmarchelier, C., Nowicki, M. & Bott, R. A Combination of Single-Nucleotide
725 Polymorphisms Is Associated with Interindividual Variability in Dietary beta-Carotene
726 Bioavailability in Healthy Men. *J Nutr* **145**, 1740-7 (2015).
- 727 16. von Lintig, J. & Wyss, A. Molecular analysis of vitamin A formation: cloning and
728 characterization of beta-carotene 15,15'-dioxygenases. *Arch Biochem Biophys* **385**, 47-52 (2001).
- 729 17. Meyers, K.J. *et al.* Genetic evidence for role of carotenoids in age-related macular degeneration
730 in the Carotenoids in Age-Related Eye Disease Study (CAREDS). *Invest Ophthalmol Vis Sci* **55**,
731 587-99 (2014).
- 732 18. Mrowicka, M., Mrowicki, J., Kucharska, E. & Majsterek, I. Lutein and Zeaxanthin and Their
733 Roles in Age-Related Macular Degeneration-Neurodegenerative Disease. *Nutrients* **14**(2022).
- 734 19. Chew, E.Y. *et al.* Long-term Outcomes of Adding Lutein/Zeaxanthin and omega-3 Fatty Acids to
735 the AREDS Supplements on Age-Related Macular Degeneration Progression: AREDS2 Report
736 28. *JAMA Ophthalmol* **140**, 692-698 (2022).

- 737 20. Moran, N.E. *et al.* Single Nucleotide Polymorphisms in beta-Carotene Oxygenase 1 are
738 Associated with Plasma Lycopene Responses to a Tomato-Soy Juice Intervention in Men with
739 Prostate Cancer. *J Nutr* **149**, 381-397 (2019).
- 740 21. Mondul, A.M. *et al.* Genome-wide association study of circulating retinol levels. *Hum Mol Genet*
741 **20**, 4724-31 (2011).
- 742 22. Zubair, N. *et al.* Genetic variation predicts serum lycopene concentrations in a multiethnic
743 population of postmenopausal women. *J Nutr* **145**, 187-92 (2015).
- 744 23. Cruz, L.A., Cooke Bailey, J.N. & Crawford, D.C. Importance of Diversity in Precision Medicine:
745 Generalizability of Genetic Associations Across Ancestry Groups Toward Better Identification of
746 Disease Susceptibility Variants. *Annu Rev Biomed Data Sci* **6**, 339-356 (2023).
- 747 24. Uffelmann, E., Posthuma, D. & Peyrot, W.J. Genome-wide association studies of polygenic risk
748 score-derived phenotypes may lead to inflated false positive rates. *Sci Rep* **13**, 4219 (2023).
- 749 25. George, S.H.L., Medina-Rivera, A., Idaghdour, Y., Lappalainen, T. & Gallego Romero, I.
750 Increasing diversity of functional genetics studies to advance biological discovery and human
751 health. *Am J Hum Genet* **110**, 1996-2002 (2023).
- 752 26. Jilcott Pitts, S. *et al.* Reflection Spectroscopy-Assessed Skin Carotenoids Are Sensitive to Change
753 in Carotenoid Intake in a 6-Week Randomized Controlled Feeding Trial in a Racially/Ethnically
754 Diverse Sample. *J Nutr* **153**, 1133-1142 (2023).
- 755 27. Jilcott Pitts, S.B. *et al.* Pressure-Mediated Reflection Spectroscopy Criterion Validity as a
756 Biomarker of Fruit and Vegetable Intake: A 2-Site Cross-Sectional Study of 4 Racial or Ethnic
757 Groups. *J Nutr* **152**, 107-116 (2022).
- 758 28. Jilcott Pitts, S.B. *et al.* A non-invasive assessment of skin carotenoid status through reflection
759 spectroscopy is a feasible, reliable and potentially valid measure of fruit and vegetable
760 consumption in a diverse community sample. *Public Health Nutr* **21**, 1664-1670 (2018).
- 761 29. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74
762 (2015).
- 763 30. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-
764 1287 (2016).
- 765 31. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for
766 disease from genome-wide association studies. *Am J Hum Genet* **88**, 294-305 (2011).
- 767 32. Tanaka, S. *et al.* C3G, a guanine nucleotide-releasing protein expressed ubiquitously, binds to the
768 Src homology 3 domains of CRK and GRB2/ASH proteins. *Proc Natl Acad Sci U S A* **91**, 3443-7
769 (1994).
- 770 33. Takai, S. *et al.* Mapping of the human C3G gene coding a guanine nucleotide releasing protein
771 for Ras family to 9q34.3 by fluorescence in situ hybridization. *Hum Genet* **94**, 549-50 (1994).
- 772 34. Gutierrez-Uzquiza, A. *et al.* C3G down-regulates p38 MAPK activity in response to stress by
773 Rap-1 independent mechanisms: involvement in cell death. *Cell Signal* **22**, 533-42 (2010).
- 774 35. Chavkin, N.W. *et al.* Adapter Protein RapGEF1 Is Required for ERK1/2 Signaling in Response to
775 Elevated Phosphate in Vascular Smooth Muscle Cells. *J Vasc Res* **58**, 277-285 (2021).
- 776 36. Vishnu, V.V. *et al.* C3G Regulates STAT3, ERK, Adhesion Signaling, and Is Essential for
777 Differentiation of Embryonic Stem Cells. *Stem Cell Rev Rep* **17**, 1465-1477 (2021).
- 778 37. Okbay, A. *et al.* Polygenic prediction of educational attainment within and between families from
779 genome-wide association analyses in 3 million individuals. *Nat Genet* **54**, 437-449 (2022).
- 780 38. Lee, J.J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of
781 educational attainment in 1.1 million individuals. *Nat Genet* **50**, 1112-1121 (2018).
- 782 39. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
- 783 40. Jeronimo, C. & Robert, F. The Mediator Complex: At the Nexus of RNA Polymerase II
784 Transcription. *Trends Cell Biol* **27**, 765-783 (2017).
- 785 41. Rachez, C. *et al.* Ligand-dependent transcription activation by nuclear receptors requires the
786 DRIP complex. *Nature* **398**, 824-8 (1999).

- 787 42. Ito, M. *et al.* Identity between TRAP and SMCC complexes indicates novel pathways for the
788 function of nuclear receptors and diverse mammalian activators. *Mol Cell* **3**, 361-70 (1999).
- 789 43. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide
790 association scans. *Bioinformatics* **26**, 2190-1 (2010).
- 791 44. Zhang, C., Hansen, M.E.B. & Tishkoff, S.A. Advances in integrative African genomics. *Trends*
792 *Genet* **38**, 152-168 (2022).
- 793 45. Martin, A.R. *et al.* Low-coverage sequencing cost-effectively detects known and novel variation
794 in underrepresented populations. *Am J Hum Genet* **108**, 656-668 (2021).
- 795 46. Bohn, T. *et al.* beta-Carotene in the human body: metabolic bioactivation pathways - from
796 digestion to tissue distribution and excretion. *Proc Nutr Soc* **78**, 68-87 (2019).
- 797 47. Borel, P. Genetic variations involved in interindividual variability in carotenoid status. *Mol Nutr*
798 *Food Res* **56**, 228-40 (2012).
- 799 48. Dron, J.S. *et al.* Association of Rare Protein-Truncating DNA Variants in APOB or PCSK9 With
800 Low-density Lipoprotein Cholesterol Level and Risk of Coronary Heart Disease. *JAMA Cardiol*
801 **8**, 258-267 (2023).
- 802 49. Vattikuti, S., Guo, J. & Chow, C.C. Heritability and genetic correlations explained by common
803 SNPs for metabolic syndrome traits. *PLoS Genet* **8**, e1002637 (2012).
- 804 50. van Dongen, J., Willemsen, G., Chen, W.M., de Geus, E.J. & Boomsma, D.I. Heritability of
805 metabolic syndrome traits in a large population-based sample. *J Lipid Res* **54**, 2914-23 (2013).
- 806 51. Watanabe, T. *et al.* Identification of immunoglobulin superfamily 11 (IGSF11) as a novel target
807 for cancer immunotherapy of gastrointestinal and hepatocellular carcinomas. *Cancer Sci* **96**, 498-
808 506 (2005).
- 809 52. Hayano, Y. *et al.* IgSF11 homophilic adhesion proteins promote layer-specific synaptic assembly
810 of the cortical interneuron subtype. *Sci Adv* **7**(2021).
- 811 53. Harada, H., Suzu, S., Hayashi, Y. & Okada, S. BT-IgSF, a novel immunoglobulin superfamily
812 protein, functions as a cell adhesion molecule. *J Cell Physiol* **204**, 919-26 (2005).
- 813 54. Scepanovic, P. *et al.* A comprehensive assessment of demographic, environmental, and host
814 genetic associations with gut microbiome diversity in healthy individuals. *Microbiome* **7**, 130
815 (2019).
- 816 55. Cheng, B. *et al.* Gut microbiota is associated with bone mineral density : an observational and
817 genome-wide environmental interaction analysis in the UK Biobank cohort. *Bone Joint Res* **10**,
818 734-741 (2021).
- 819 56. Pulit, S.L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694
820 649 individuals of European ancestry. *Hum Mol Genet* **28**, 166-174 (2019).
- 821 57. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am J*
822 *Hum Genet* **104**, 65-75 (2019).
- 823 58. Zhu, Z. *et al.* Shared genetic and experimental links between obesity-related traits and asthma
824 subtypes in UK Biobank. *J Allergy Clin Immunol* **145**, 537-549 (2020).
- 825 59. Lim, J.Y. & Wang, X.D. Mechanistic understanding of beta-cryptoxanthin and lycopene in cancer
826 prevention in animal models. *Biochim Biophys Acta Mol Cell Biol Lipids* **1865**, 158652 (2020).
- 827 60. Burri, B.J., La Frano, M.R. & Zhu, C. Absorption, metabolism, and functions of beta-
828 cryptoxanthin. *Nutr Rev* **74**, 69-82 (2016).
- 829 61. Burri, B.J. Beta-cryptoxanthin as a source of vitamin A. *J Sci Food Agric* **95**, 1786-94 (2015).
- 830 62. Lai, J.S. *et al.* Higher maternal plasma beta-cryptoxanthin concentration is associated with better
831 cognitive and motor development in offspring at 2 years of age. *Eur J Nutr* **60**, 703-714 (2021).
- 832 63. Grassmann, S. *et al.* SNP rs6564851 in the BCO1 Gene Is Associated with Varying Provitamin A
833 Plasma Concentrations but Not with Retinol Concentrations among Adolescents from Rural
834 Ghana. *Nutrients* **12**(2020).
- 835 64. Yang, J.S. *et al.* ALDH7A1 inhibits the intracellular transport pathways during hypoxia and
836 starvation to promote cellular energy homeostasis. *Nat Commun* **10**, 4068 (2019).

- 837 65. Tam, A.B. *et al.* The UPR Activator ATF6 Responds to Proteotoxic and Lipotoxic Stress by
838 Distinct Mechanisms. *Dev Cell* **46**, 327-343 e7 (2018).
- 839 66. Sekine, H. *et al.* The Mediator Subunit MED16 Transduces NRF2-Activating Signals into
840 Antioxidant Gene Expression. *Mol Cell Biol* **36**, 407-20 (2016).
- 841 67. Fixsen, B.R. *et al.* SALL1 enforces microglia-specific DNA binding and function of SMADs to
842 establish microglia identity. *Nat Immunol* **24**, 1188-1199 (2023).
- 843 68. Baumann, C.A. *et al.* CAP defines a second signalling pathway required for insulin-stimulated
844 glucose transport. *Nature* **407**, 202-7 (2000).
- 845 69. Liu, M.M. *et al.* SORBS2 as a molecular target for atherosclerosis in patients with familial
846 hypercholesterolemia. *J Transl Med* **20**, 233 (2022).
- 847 70. Moncan, M. *et al.* Regulation of lipid metabolism by the unfolded protein response. *J Cell Mol*
848 *Med* **25**, 1359-1370 (2021).
- 849 71. Momozawa, Y. *et al.* Low-frequency coding variants in CETP and CFB are associated with
850 susceptibility of exudative age-related macular degeneration in the Japanese population. *Hum Mol*
851 *Genet* **25**, 5027-5034 (2016).
- 852 72. Cheng, C.Y. *et al.* New loci and coding variants confer risk for age-related macular degeneration
853 in East Asians. *Nat Commun* **6**, 6063 (2015).
- 854 73. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height.
855 *Nature* **610**, 704-712 (2022).
- 856 74. Thornton, T.A. & Bermejo, J.L. Local and global ancestry inference and applications to genetic
857 association analysis for admixed populations. *Genet Epidemiol* **38 Suppl 1**, S5-S12 (2014).
- 858 75. Ermakov, I.V. & Gellermann, W. Dermal carotenoid measurements via pressure mediated
859 reflection spectroscopy. *J Biophotonics* **5**, 559-70 (2012).
- 860 76. Qiang Wu *et al.* A reflection-spectroscopy measured skin carotenoid score strongly correlates
861 with plasma concentrations of all major dietary carotenoid species except for lycopene. *Nutrition*
862 *Research* (2024).
- 863 77. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies.
864 *Nat Genet* **44**, 821-4 (2012).
- 865 78. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
866 *Bioinformatics* **25**, 1754-60 (2009).
- 867 79. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**(2021).
- 868 80. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic
869 features. *Bioinformatics* **26**, 841-2 (2010).
- 870 81. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-
871 generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).
- 872 82. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation
873 DNA sequencing data. *Nat Genet* **43**, 491-8 (2011).
- 874 83. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome
875 Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 1-11 10 33 (2013).
- 876 84. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from
877 high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
- 878 85. Wang, G.T., Peng, B. & Leal, S.M. Variant association tools for quality control and analysis of
879 large-scale sequence and genotyping array data. *Am J Hum Genet* **94**, 770-83 (2014).
- 880 86. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-8 (2011).
- 881
- 882

883

FIGURES and TABLES

884

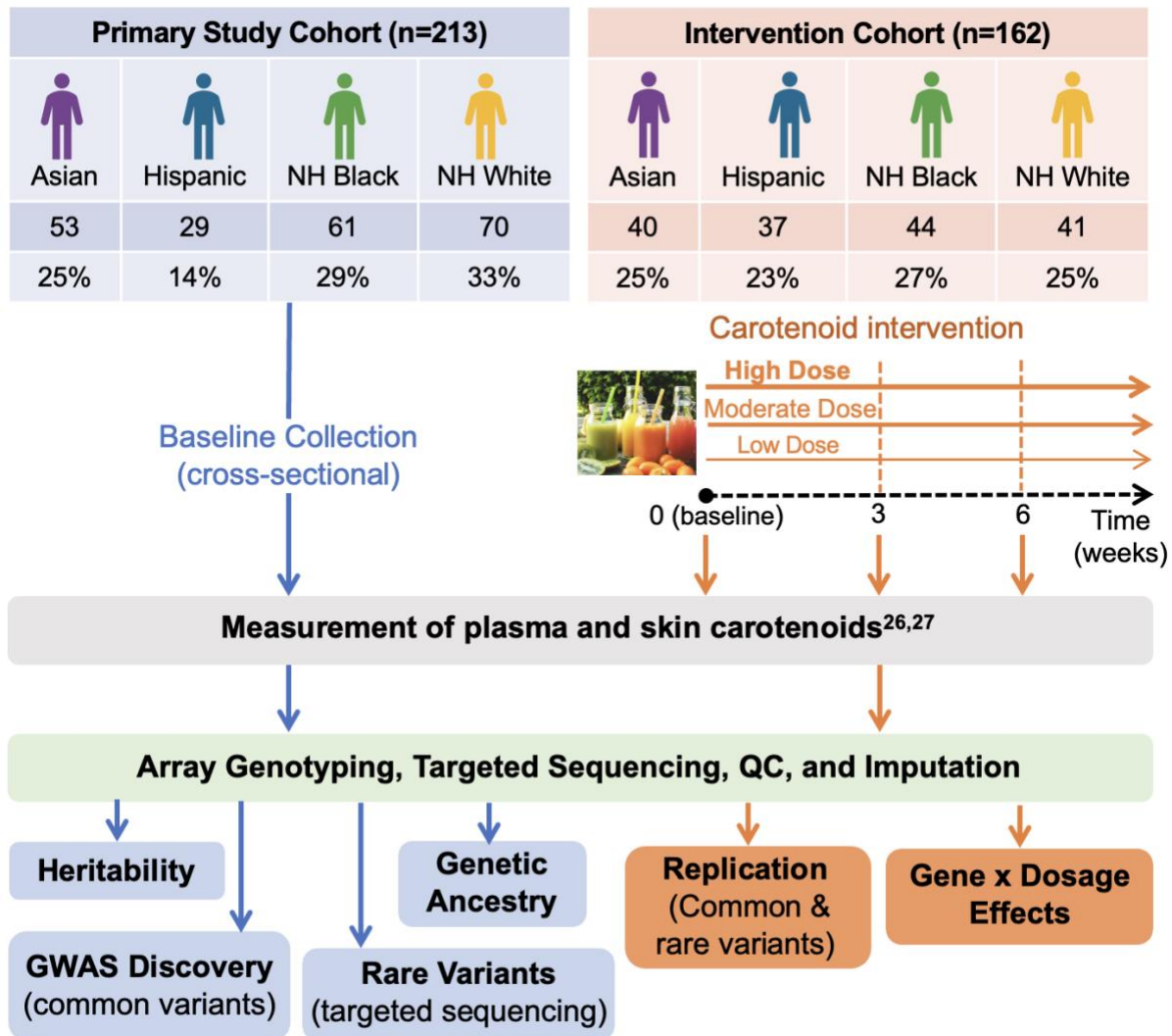
Common and rare genetic variation intersects with ancestry to

885

influence human skin and plasma carotenoid concentrations

886

(Han et. al)

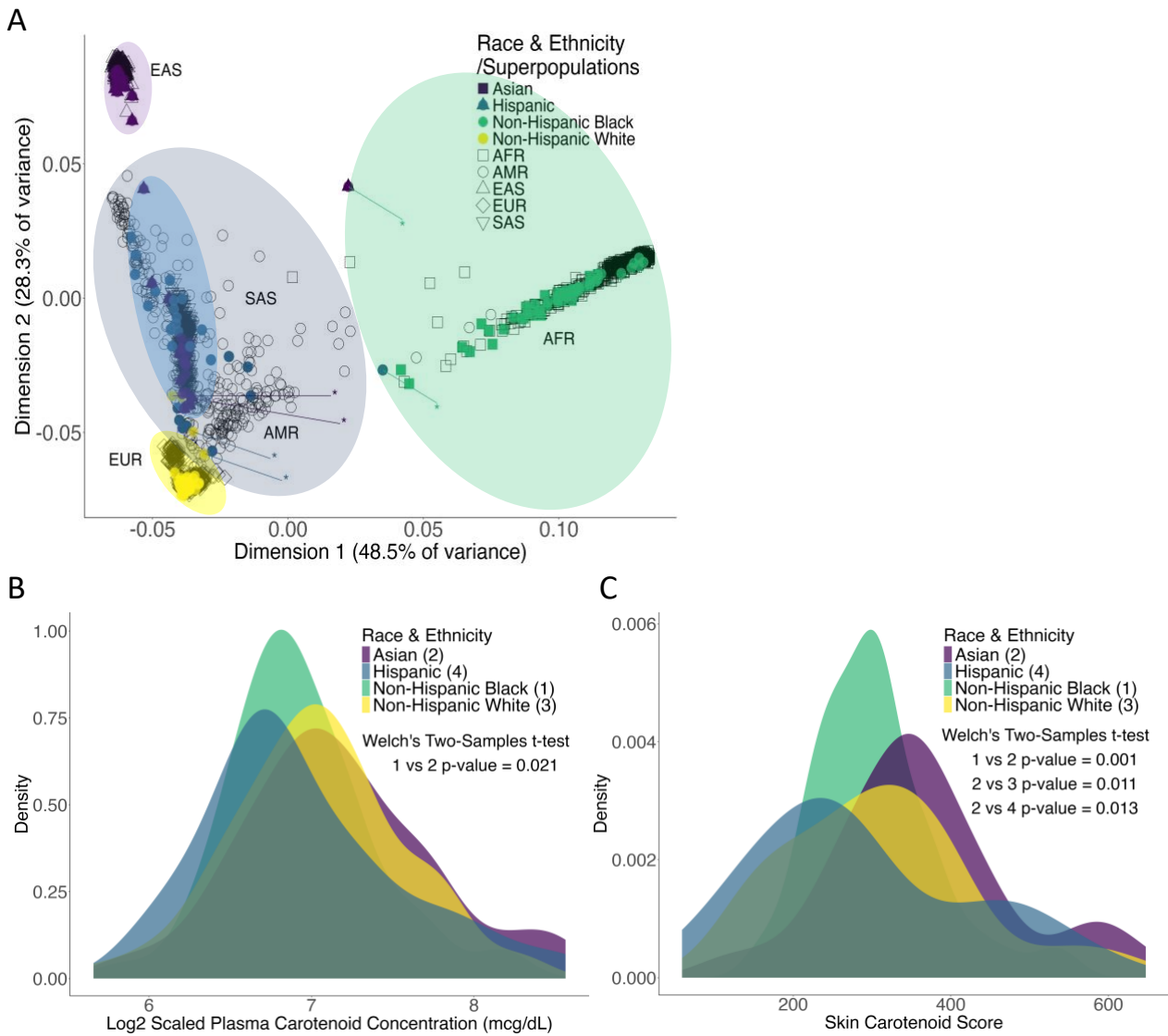


887

888 **Figure 1. Overview of study design and analysis workflow.** The study includes participants
 889 from two cohorts: the **Primary Study Cohort** (n=213) and the **Intervention Cohort** (n=162),
 890 both comprised of multiple racial/ethnic groups (Asian, Hispanic, non-Hispanic Black, and non-
 891 Hispanic White), with corresponding sample sizes and proportions detailed in the top panel. The
 892 Intervention Cohort was divided into three groups based on fruit and vegetable (FV) intake: Low
 893 Dose/Control (negligible carotenoid intake), Moderate Dose (4 mg total carotenoids/day), and
 894 High Dose (8 mg total carotenoids/day). Baseline measurements of plasma and skin carotenoids

895 were performed for all participants. For the Intervention Cohort, additional measurements were
896 collected at weeks 3 and 6 post-intervention. Data were collected from previous studies (see
897 Reference 6 and 8). Array Genotyping data from the Study Cohort was used for genetic ancestry
898 assessment and heritability estimation, stratified by race/ethnicity. Both common and rare
899 variants identified in the Study Cohort were further analyzed for interaction effects in the
900 Intervention Cohort.
901

902



903

904 **Figure 2. Self-reported and genetically defined race/ethnicity and total carotenoid**

905 **concentrations in primary cohort.** Self-identified race/ethnicity groups include Asian (n=53),

906 Hispanic (n=29), Non-Hispanic Black (n=61), and Non-Hispanic White (n=70). **2A** -

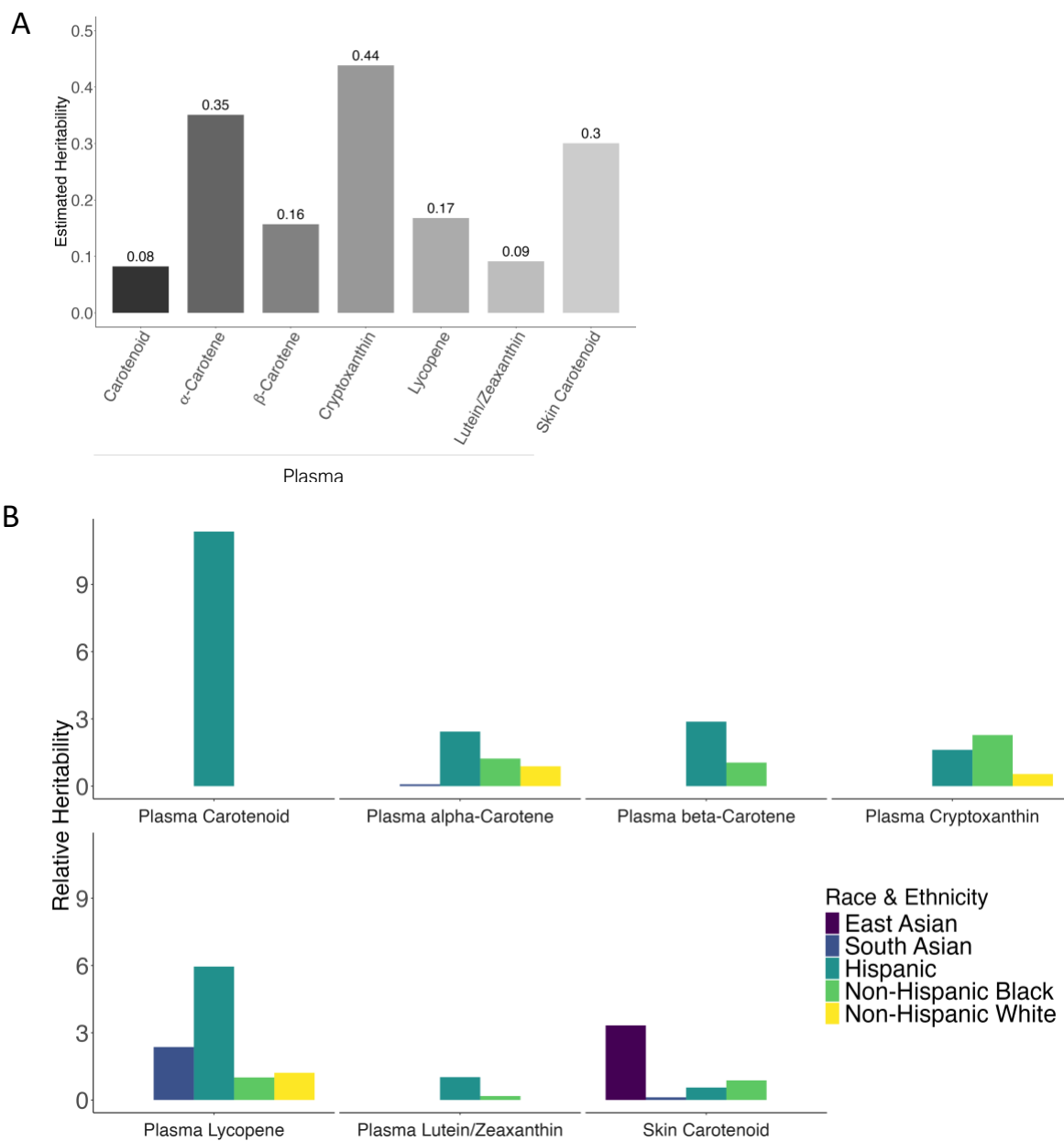
907 Multidimensional Scaling (MDS) of genomic data alongside 'superpopulations' from 1000

908 Genomes Project: AFR (African), AMR (Admixed American), EAS (East Asian), EUR

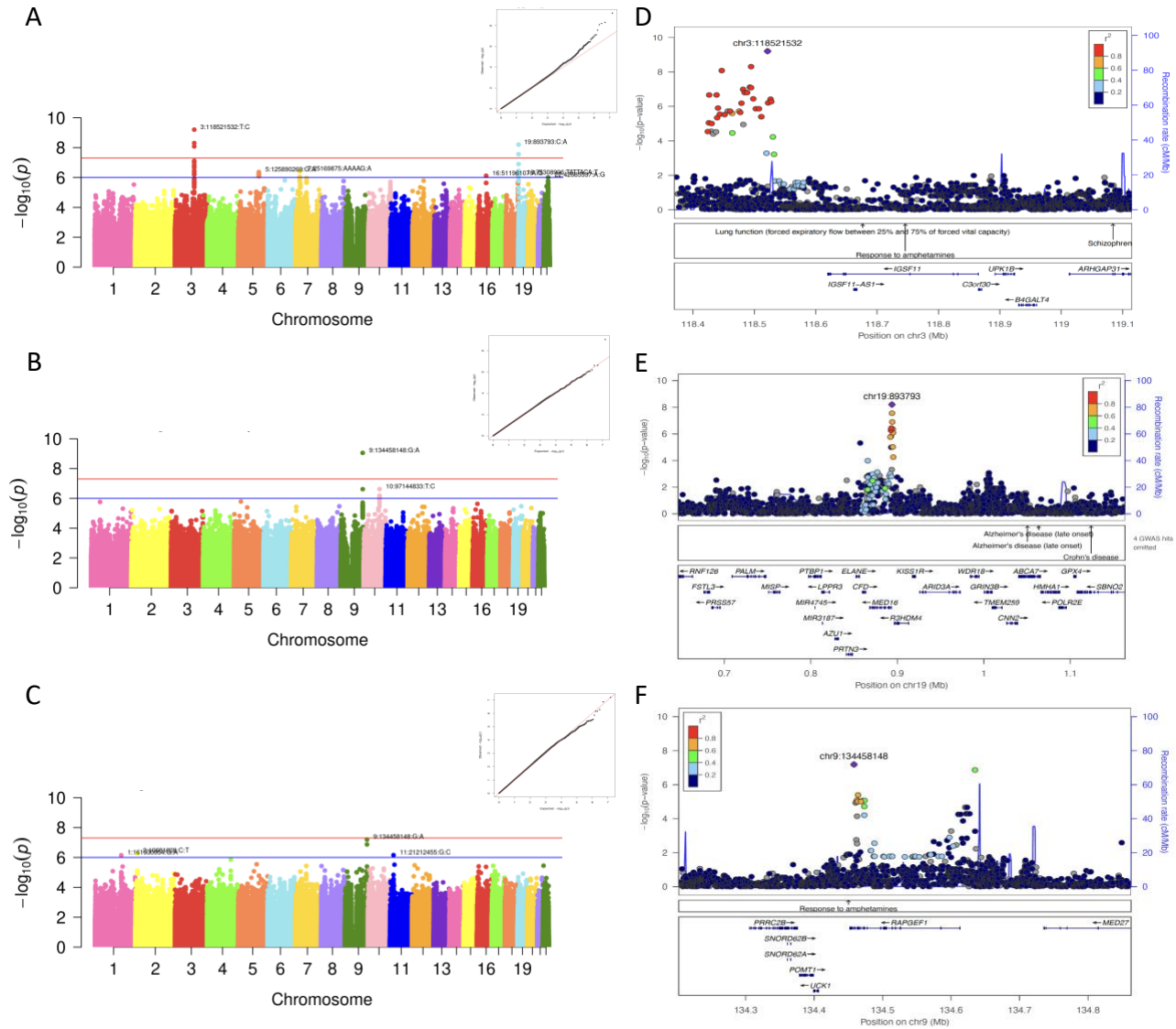
909 (European), SAS (South Asian). '*' indicates individuals whose self-reported ancestry differs

910 from genomic alignment. **2B** - Plasma carotenoid concentrations (mcg/dL, log₂-transformed); **2C**

911 - skin carotenoid scores. Group number is indicated in parentheses, and differences were
912 assessed via Welch's Two-Sample *t*-tests.



913
914 **Figure 3. Heritability of plasma and skin carotenoids. 3A** - Heritability estimates for plasma
915 carotenoid concentrations and skin carotenoid score across the entire Primary Study Cohort. **3B** -
916 Relative heritability of plasma carotenoid subspecies for race/ethnicity groups. Genetic ancestry
917 is defined based on multidimensional scaling (MDS) analysis. Relative heritability is calculated
918 as the heritability estimate in each subgroup divided by the heritability estimate for the entire
919 Primary Study Cohort for each carotenoid species.



920

921 **Figure 4. Genome-wide association analysis of plasma carotenoid and subspecies**

922 **concentrations.** Manhattan and Q-Q plots from GEMMA linear regression analysis of log₂-

923 transformed plasma cryptoxanthin (4A), α -carotene (4B), and β -carotene (4C). Genome-wide

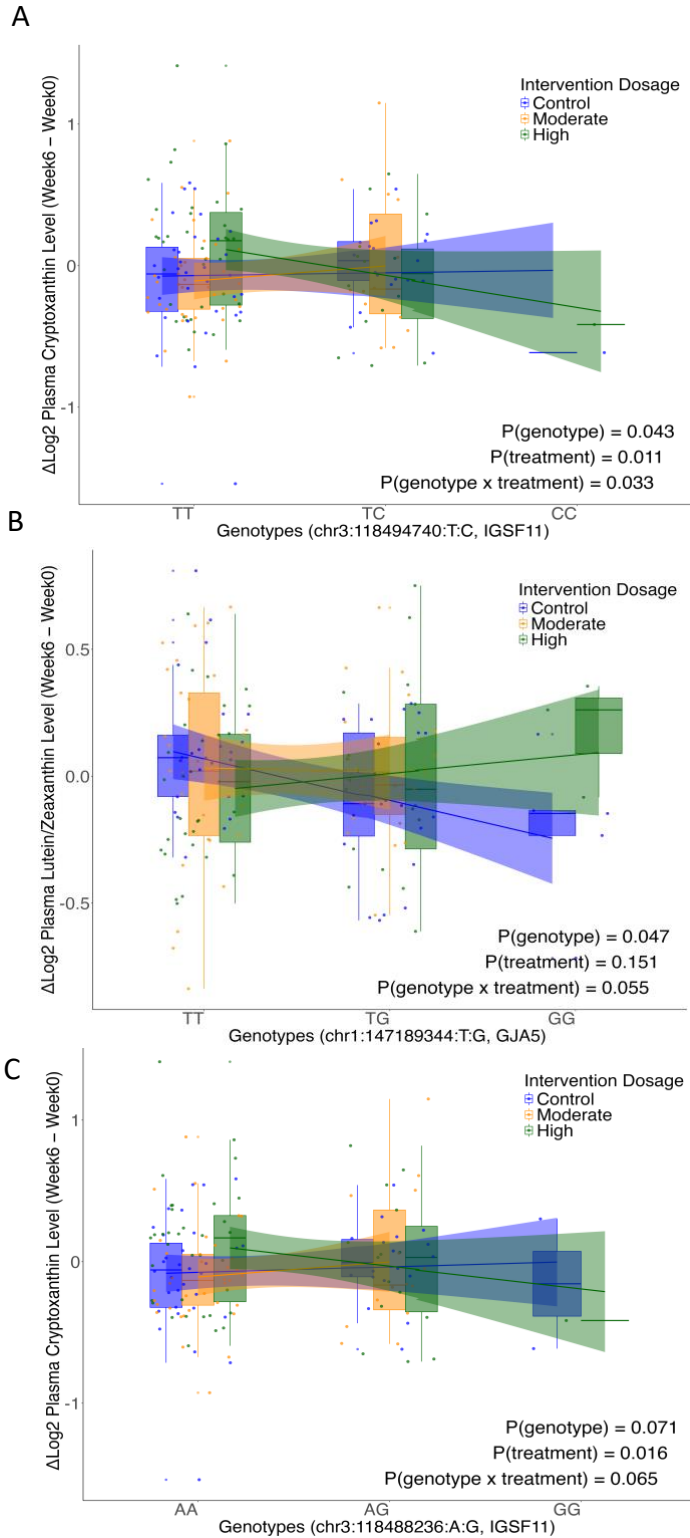
924 significance is indicated by red line ($-\log P = 7.3$) and suggestive association by the blue line ($-\log P = 6$).

925 LocusZoom plots of SNP associations with carotenoid levels at significant loci: 3q13

926 (*IGSF1*) (4D), 19p13 (*R3HDM4/MED16*) (4E), and intragenic to *RAPGEF1* (4F).

927 Recombination rates and linkage disequilibrium (r^2) are relative to the AFR superpopulation in

928 1000 Genomes.

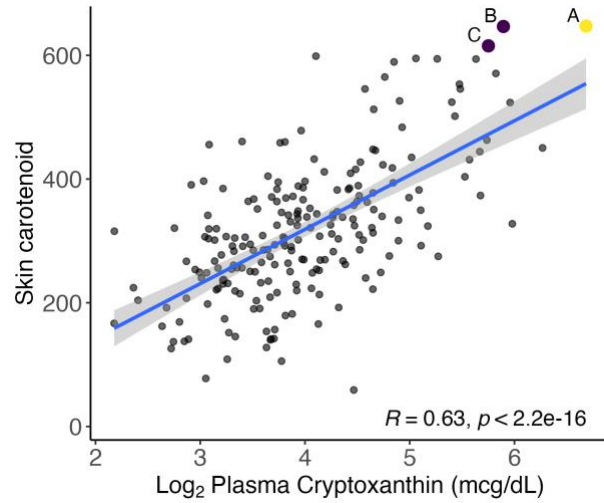


929

930 **Figure 5. Gene-by-dosage plots.** Genotypes are shown along the x-axis, with intervention

931 dosages color-coded: blue - Control, orange - Moderate, and green - High. Box plots represent

932 plasma carotenoid distributions within each genotype and dosage. Smoothed regression lines
933 show genotype effects within treatments. P-values indicate the significance of genotype
934 (P(genotype)), treatment (P(treatment)), and genotype-by-treatment interaction (P(genotype ×
935 treatment)). Treatment p-values < 0.1 indicate significant dosage effects.



936

937 **Figure 6. Correlation between skin carotenoid levels and plasma cryptoxanthin.** Scatterplot

938 shows skin carotenoid levels (y-axis) and log₂-transformed plasma cryptoxanthin concentrations

939 (x-axis). Each gray dot represents a participant. The blue line represents the linear regression fit,

940 with a significant positive correlation ($R = 0.63$, $p < 2.2 \times 10^{-16}$). Outliers (purple; B and C), who

941 self-identified and clustered by genomic information as Asian, carry the rare East Asian-specific

942 *CETP* variant (rs2303790). Outlier (yellow; A), self-identified and clustered by genomic

943 information as Non-Hispanic White, has a rare *APOA1* variant (rs756535387) predicted to be

944 deleterious. Details of deleterious variants are given in **Supplementary Table S9**.

945

946 **TABLES**

Phenotype (Plasma)	Nearest Gene	SNPs	rsID	Primary Study Cohort			Intervention Cohort	
				AF	Beta	P	P _{Meta}	Dir
α-Carotene	RAPGEF1	9:134458148:G:A	rs3765544	0.054	0.75	8.86E-10	1.74E-09	+?
β-Carotene	RAPGEF1	9:134458148:G:A	rs3765544	0.054	0.79	6.43E-08	1.01E-07	+?
Cryptoxanthin	<i>IGSF11</i>	3:118446886:T:C	rs76087842	0.08	0.72	8.32E-09	1.32E-06	--
Cryptoxanthin	<i>IGSF11</i>	3:118492749:G:A	rs798600	0.129	0.55	7.63E-08	3.52E-06	--
Cryptoxanthin	<i>IGSF11</i>	3:118494728:C:G	rs76613159	0.078	0.71	4.95E-09	1.62E-08	--
Cryptoxanthin	<i>IGSF11</i>	3:118494740:T:C	rs1088589	0.137	0.53	8.21E-08	1.08E-04	+-
Cryptoxanthin	<i>IGSF11</i>	3:118521532:T:C	rs80316816	0.08	0.74	6.25E-10	1.34E-06	+-
Cryptoxanthin	<i>R3HDM4</i>	19:893793:C:A	rs28468554	0.359	-0.46	6.26E-09	1.76E-06	--
Cryptoxanthin	<i>R3HDM4</i>	19:893910:A:G	rs28626160	0.495	-0.37	2.77E-08	1.96E-05	+-

947 **Table 1. SNPs significantly associated with carotenoid concentration phenotypes.** SNPs in
 948 bold were directly genotyped on the Infinium™ H3Africa Consortium Array v2; the remaining
 949 SNPs were imputed. SNP ID includes chromosome: chromosome position: reference allele:
 950 alternate allele. Phenotypes are log₂-transformed carotenoid concentrations in plasma. AF-Minor
 951 Allele Frequency; Beta—logistic regression slope (effect size); P - p-value from the GEMMA
 952 analysis; P_{Meta} - p-value from the METALysis. Dir – direction of effect (positive or negative).
 953