

# 1 Limited predictability of amino acid substitutions in seasonal influenza viruses

2 Pierre Barrat-Charlaix,<sup>1,2</sup> John Huddleston,<sup>3</sup> Trevor Bedford,<sup>4</sup> and Richard A. Neher<sup>1,2,\*</sup>

3 <sup>1</sup>*Biozentrum, Universität Basel,*

4 *Switzerland*

5 <sup>2</sup>*Swiss Institute of Bioinformatics, Basel,*

6 *Switzerland*

7 <sup>3</sup>*Molecular Cell Biology,*

8 *University of Washington, Seattle, WA,*

9 *USA*

10 <sup>4</sup>*Vaccine and Infectious Disease Division,*

11 *Fred Hutchinson Cancer Research Center, Seattle, WA,*

12 *USA*

13 (Dated: December 11, 2020)

14 Seasonal influenza viruses repeatedly infect humans in part because they rapidly change  
15 their antigenic properties and evade host immune responses, necessitating frequent  
16 updates of the vaccine composition. Accurate predictions of strains circulating in the  
17 future could therefore improve the vaccine match. Here, we studied the predictability of  
18 frequency dynamics and fixation of amino acid substitutions. Current frequency was the  
19 strongest predictor of eventual fixation, as expected in neutral evolution. Other properties,  
20 such as occurrence in previously characterized epitopes or high *Local Branching Index*  
21 (LBI) had little predictive power. Parallel evolution was found to be moderately predictive  
22 of fixation. While the LBI had little power to predict frequency dynamics, it was still  
23 successful at picking strains representative of future populations. The latter is due to a  
24 tendency of the LBI to be high for consensus-like sequences that are closer to the future  
25 than the average sequence. Simulations of models of adapting populations, in contrast,  
26 show clear signals of predictability. This indicates that the evolution of influenza HA and  
27 NA, while driven by strong selection pressure to change, is poorly described by common  
28 models of directional selection such as travelling fitness waves.

## 29 INTRODUCTION

30 Seasonal influenza A viruses (IAV) infect about 10% of 53  
31 the global population every year, resulting in hundreds 54  
32 of thousands of deaths (Organization, 2018; Petrova and 55  
33 Russell, 2017). Vaccination is the primary measure to 56  
34 reduce influenza morbidity. However, the surface proteins 57  
35 hemagglutinin (HA) and neuraminidase (NA) continu- 58  
36 ously accumulate mutations at a high rate, leading to fre- 59  
37 quent antigenic changes (Petrova and Russell, 2017; Shih 60  
38 et al., 2007; Bhatt et al., 2011; Koel et al., 2013). While 61  
39 a vaccine targeting a particular strain may be efficient for 62  
40 some time, antigenic drift will sooner or later render it ob- 63  
41 solete. The World Health Organization (WHO) regularly 64  
42 updates influenza vaccine recommendations to best match 65  
43 the circulating strains. Since developing, manufacturing, 66  
44 and distributing the vaccine takes many months, forecast- 67  
45 ing the evolution of influenza is of essential interest to 68  
46 public health (Morris et al., 2018; Klingen et al., 2018a). 69

47 The number of available high quality HA and 70  
48 NA sequences has increased rapidly over the last 20 71  
49 years (Bogner et al., 2006; Shu and McCauley, 2017) 72  
50 and virus evolution and dynamics can be now be tracked 73

51 at high temporal and spatial resolution (Rambaut et al.,  
52 2008). This wealth of data has given rise to an active  
53 field of predicting influenza virus evolution (Morris et al.,  
54 2018; Klingen et al., 2018a). These models predict the  
55 future population of influenza viruses by estimating strain  
56 fitness or proxies of fitness. Luksza and Lässig (2014),  
57 for example, train a fitness model to capture antigenic  
58 drift and protein stability on patterns of epitope and non-  
59 epitope mutations. Other approaches by Steinbrück et al.  
60 (2014); Neher et al. (2016) predict fitness by using hemag-  
61 glutination inhibition (HI) data to determine possible  
62 antigenic drift of clades in the genealogy of the HA pro-  
63 tein. Finally, Neher et al. (2014) use branching patterns  
64 of HA phylogenies as a proxy for fitness. These branching  
65 patterns are summarized by the Local Branching Index  
66 (LBI), which was shown to be a proxy of relative fitness  
67 in mathematical models of rapidly adapting populations  
68 (Neher et al., 2014).

69 The underlying assumption of all these methods is that  
70 (i) differences in growth rate between strains can be es-  
71 timated from sequence or antigenic data and (ii) that  
72 these growth rate differences persist for long enough to be  
73 predictive of future success. Specific positions in surface  
74 proteins are of particular interest in this context. The  
75 surface proteins are under a strong positive selection and  
76 change their amino acid sequence much more rapidly than  
77 other IAV proteins or than expected under neutral evo-  
78 lution (Bhatt et al., 2011; Strelkowa and Lässig, 2012).

\* Correspondence to: Richard Neher, Biozentrum, 76  
Klingelbergstrasse 70, 4056, Basel, Switzerland. 77  
richard.neher@unibas.ch 78

Epitope positions, i.e., positions targeted by human antibodies, are expected to change particularly often since viruses with altered epitopes can evade existing immune responses (Shih et al., 2007; Koel et al., 2013; Wolf et al., 2006). It therefore seems plausible that mutations at these positions have a tendency to increase fitness and a higher probability of fixation (Strelkowa and Lässig, 2012). But one has to be careful to account for the fact that these positions are often ascertained post-hoc (Shih et al., 2007) and human immune responses are diverse with substantial inter-individual variation (Lee et al., 2019).

A mutation with a fitness benefit should have a tendency to fix, while detrimental mutations fix less often

on the same segment, and the effect of the remaining seven segments. Here, we investigate properties of broad categories of mutations effectively averaging over different genetic backgrounds to isolate the effects intrinsic to the mutation.

First, we ask whether we can quantitatively predict the frequency of  $X$  at future times  $f(t)$ . In other words, having observed a mutation at frequencies  $(f_1, f_2, \dots, f_n)$  at dates  $(t_1, t_2, \dots, t_n)$ , what can we say about its frequency at future dates  $(t_{n+1}, t_{n+2}, \dots)$ ? A simpler, more qualitative question, is to ask whether  $X$  will fix in the population, will disappear, or whether the site will stay polymorphic.

We use amino-acid sequences of the HA and NA genes of A/H3N2 (since the year 2000) and A/H1N1pdm (since the year 2009) influenza available in GISAID (Shu and McCauley, 2017) (see supplementary materials for an acknowledgment of all data contributors). This amounts to 44 976 HA and 36 300 NA sequences for A/H3N2 and 45 350 HA and 40 412 NA sequences for A/H1N1, with a minimum of 100 per year. These sequences are binned in non-overlapping intervals of one month. Each single-month time bin and the sequences that it contains represent a (noisy) snapshot of the influenza population at a given date. The number of sequences per time bin varies strongly both with year and according to the season, with earlier time bins containing around 10 sequences while more recent bins contain several hundreds (see figures S9 and S10 in SM for details).

The central quantities that we derived from this data are *frequency trajectories* of amino acids at each position in the sequences. If an amino acid  $X_i$  is found at position  $i$  at a frequency between 5% and 95% in the population of a given time bin  $t$ , then the population is considered polymorphic at position  $i$  and at time  $t$ . This polymorphism is characterized by the frequency of  $X_i$ ,  $f_{X_i}(t)$ , and also by frequencies of other amino acids at  $i$ . The series of values  $f_{X_i}(t)$  for contiguous time bins constitutes the frequency trajectory of  $X_i$ . A trajectory is terminated if the corresponding frequency is measured above 95% (resp. below 5%) for two time bins in a row, in which case amino acid  $X_i$  is considered as *fixed* (resp. *absent*) in the population. Otherwise, the trajectory is considered *active*. Examples of trajectories can be seen in figure S11 of the Supplement.

Since genomic segments of influenza do not recombine, concurrent trajectories of mutations on the same segment will be correlated. In the extreme case were the two mutations always appear on the same strains, their two trajectories will be identical. This “nested” character or frequency trajectories may introduce biases when performing statistical analysis. In section 3 of the SM, we investigate the effect of such nested trajectories by clustering them based on the strains that compose them. While the number of trajectories that contribute to the analysis decreases with more aggressive clustering, this

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

<sup>188</sup> does not significantly change the results presented below<sup>241</sup>

<sup>189</sup> the sense that they start at frequency 0 for  $t \rightarrow -\infty$ .  
<sup>190</sup> Removing them does not change results significantly.

<sup>191</sup> In the rest of this work, we will focus on frequency  
<sup>192</sup> trajectories that are starting at a zero (low) frequency,  
<sup>193</sup> i.e.  $f(t = 0) = 0$ . These represent new amino acid vari-  
<sup>194</sup> ants which were absent in the population at the time bin  
<sup>195</sup> when the trajectory started and are currently rising in the  
<sup>196</sup> population (see Methods). Such distinction in novel and  
<sup>197</sup> ancestral variants is necessary to meaningfully interrogate  
<sup>198</sup> predictability. Each rising trajectory of a new mutation  
<sup>199</sup> implies the existence of another decreasing one at the  
<sup>200</sup> same position, since frequencies of all amino acids at a<sup>251</sup>  
<sup>201</sup> given position must sum to one. If novel variants arise by<sup>252</sup>  
<sup>202</sup> selection, we expect to see a stronger signal of selection<sup>253</sup>  
<sup>203</sup> after conditioning on these novel variants. In classic mod-<sup>254</sup>  
<sup>204</sup> els of population genetics, strongly advantageous variants<sup>255</sup>  
<sup>205</sup> undergo rapid selective *sweeps*, i.e., the rapid rise and<sup>256</sup>  
<sup>206</sup> fixation. The sweep of a mutation can be due to its own<sup>257</sup>  
<sup>207</sup> fitness effect, to the genetic background or to the effect of<sup>258</sup>  
<sup>208</sup> the seven other segments. By considering the ensemble of<sup>259</sup>  
<sup>209</sup> novel variants that are rising in frequency, we effectively<sup>260</sup>  
<sup>210</sup> average over backgrounds, obtaining a set of mutations<sup>261</sup>  
<sup>211</sup> that we expect to be beneficial on average. If such sweeps<sup>262</sup>  
<sup>212</sup> are common in the evolution of HA and NA, the restric-<sup>263</sup>  
<sup>213</sup> tion to trajectories that start at low frequency should<sup>264</sup>  
<sup>214</sup> thus enrich for mutations that are positively selected and<sup>265</sup>  
<sup>215</sup> on their way to fixation.

## <sup>215</sup> Predicting future frequencies

<sup>216</sup> Having observed the frequency trajectory  $f(t)$  of a<sup>271</sup>  
<sup>217</sup> mutation until a given date  $t_0$ , how much can we say<sup>272</sup>  
<sup>218</sup> about the future values of  $f$  after  $t_0$ ? We consider the<sup>273</sup>  
<sup>219</sup> idealized case sketched in panel **A** of figure 1: given<sup>274</sup>  
<sup>220</sup> the trajectory of a *new* mutation, i.e. that started at a  
<sup>221</sup> frequency of 0, and that we observe at frequency  $f_0$  at<sup>275</sup>  
<sup>222</sup> time  $t_0$ , what is the probability  $P_{\Delta t}(f)$  of observing it at<sup>276</sup>  
<sup>223</sup> a value  $f$  at time  $t_0 + \Delta t$ ?<sup>277</sup>

<sup>224</sup> To answer this question retrospectively, we use all fre-<sup>278</sup>  
<sup>225</sup> quency trajectories extracted from HA and NA sequences<sup>279</sup>  
<sup>226</sup> that satisfy these conditions for a given  $f_0$ . The num-<sup>280</sup>  
<sup>227</sup> ber of trajectories is limited and the frequency estimates<sup>281</sup>  
<sup>228</sup> themselves are based on a finite sample and are hence im-<sup>282</sup>  
<sup>229</sup> precise. Therefore, we consider trajectories in an interval<sup>283</sup>  
<sup>230</sup>  $[f_0 - \delta f, f_0 + \delta f]$  with  $\delta f = 0.05$ .<sup>284</sup>

<sup>231</sup> For  $f_0 = 0.3$ , we found 120 such trajectories in the<sup>285</sup>  
<sup>232</sup> case of A/H3N2 influenza, represented on the panel **B**<sup>286</sup>  
<sup>233</sup> of figure 1, where time is shifted such that  $t_0 = 0$ . The  
<sup>234</sup> same analysis was performed for A/H1N1pdm, with the<sup>288</sup>  
<sup>235</sup> 89 found trajectories displayed in figure S13. Some trajec-<sup>289</sup>  
<sup>236</sup> tories fall in the frequency bin around  $f_0$  while decreasing,<sup>290</sup>  
<sup>237</sup> even though they crossed that bin at an earlier time. This<sup>291</sup>  
<sup>238</sup> is due to the fact that some trajectories “skipped” the<sup>292</sup>  
<sup>239</sup> interval  $f_0$  in question on their initial rise due to sparse<sup>293</sup>  
<sup>240</sup> sampling. These trajectories are nevertheless rising in<sup>294</sup>

Since rapid sequence evolution of influenza HA and NA mediates immune evasion, one could expect that a significant fraction of new amino acid mutations on rising trajectories in figure 1 are *adaptive*. We could thus expect that most of these trajectories continue to rise after reaching frequency  $f_0$ , at least for some time. A fraction of those would then sweep through the population and fix.

To quantify the extent to which this preconception of sweeping adaptive mutations is true, we estimated the probability distribution  $P_{\Delta t}(f|f_0)$  of finding a trajectory at frequency  $f$  after a time  $\Delta t$  given that it was observed at  $f_0$  at time 0. The results for different  $\Delta t$  are shown in figure 1C. Initially, i.e. at time  $t_0 = 0$ , this distribution is by construction peaked around  $f_0$ . If a large fraction of the trajectories keep increasing after this time, we should see the “mass” of  $P_{\Delta t}(f|f_0)$  move to the right towards higher frequencies as time progresses.

However, future distributions for  $\Delta t > 0$  do not seem to follow a pattern compatible with selective sweeps. The thick black line in Figure 1B shows the average frequency of all trajectories. This average makes a sharp turn at  $t = 0$  and is essentially flat for  $t > 0$  in the case of A/H3N2, and slightly increasing for A/H1N1pdm (see supplement). Hence, the fact that this average rose for  $t < 0$  gives little information for  $t > 0$ , and is due to the conditions by which these trajectories were selected. This shows that sweep-like trajectories rising steadily from frequency 0 to 1 are not common enough to dominate the average trajectory. **Pierre: This result is consistent with(?) They show that sweep-like trajectories are very bad fits to actual trajectories.**

Consistent with the average, the frequency distribution of the selected trajectories broadens in time without a significant shift of the mean as time passes. After 60 days, the distribution is rather symmetrical around the initial  $f_0 = 0.3$  value, suggesting that the knowledge that the trajectories were rising is lost after two months. On a timescale of 60 to 120 days, the only possible prediction is that trajectories are likely to be found in a broad interval around the initial frequency  $f_0$ . After one year the distribution becomes almost flat (excluding mutations that have disappeared or fixed), and the initial peak at  $f_0$  is not visible anymore. The only information remaining from the initial frequency is the fraction that fixed or was lost (see below). This behavior is expected in neutral models of evolution (Kimura, 1964) but incompatible with a dynamic dominated by sweeps taking over the population.

While this observation does not rule out that signatures exist that predict future frequency dynamics, past dynamics alone is weakly informative.

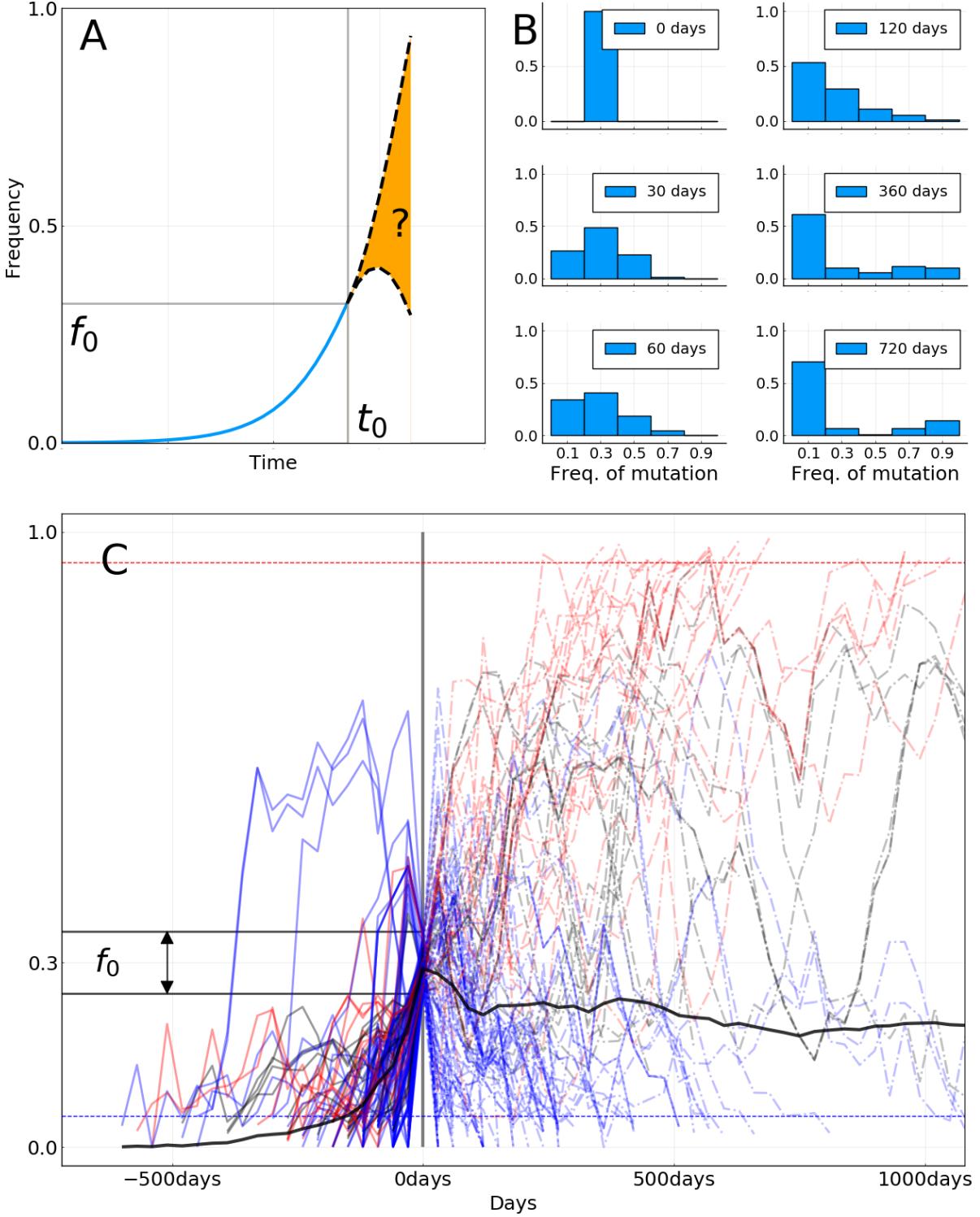


FIG. 1 **A:** Sketch of the idea behind the short term prediction of frequency trajectories. Given a mutation that we have seen increasing in frequency and that we “catch” at frequency  $f_0$  at time  $t_0$ , what can we say about the distribution of future frequencies  $P_{\Delta t}(f|f_0)$ ? **B:** All frequency trajectories of amino acid mutations in the A/H3N2 HA and NA genes that were absent in the past, are seen around  $f_0 = 30\%$  frequency at time  $t_0 = 0$ , and are based on more than 10 sequences at each time point. Red curves represent mutations that will ultimately fix, blue the ones that will be lost, and black the ones for which we do not know the final status. Dashed horizontal lines (blue and red) represent loss and fixation thresholds. The thick black line is the average of all trajectories, counting those that fix (resp. disappear) as being at frequency 1 (resp. 0). Figure S12 shows equivalent figures for other values of  $f_0$ . **C:** Distribution of future frequencies  $P_{\Delta t}(f|f_0)$  for the trajectories shown in panel **B** and for specific values of  $\Delta t$ .

## 295 Prediction of fixation or loss

350

Instead of predicting future frequency, let's consider the long-term goal of predicting the probability that a mutation fixes in the population. We first estimate the fraction of frequency trajectories that either fix in the population or are lost, as well as the time it takes for one or the other to happen. Panels **A** and **B** of figure 2 shows the fraction of frequency trajectories in HA and NA that either have fixed, were lost or remained active as a function of the time elapsed since they were first seen above 25% frequency. Most mutations are either lost or become fixed after 2-3 years, with very few trajectories remaining active after 5 years. This time scale of 2-3 years is consistent with the typical coalescence time observed in phylogenetic trees of A/H3N2 influenza (Rambaut et al., 2008; Yan et al., 2019). We also note that the fraction of lost trajectories increases sharply at small times with 40% of mutations observed above 25% frequency being lost within one year for A/H3N2, while it takes longer to fix a mutation in the whole population.

We then examined the probability of mutations to fix in the population as a function of the frequency at which they are seen. For different values of frequency  $f$ , we consider all trajectories that started at a null frequency and are seen in the interval  $[f - 7.5\%, f + 7.5\%]$  at any given time. The probability of a mutation fixing given that it is seen at frequency  $f$ ,  $P_{fix}(f)$ , is then estimated by the fraction of those trajectories which terminate at a frequency larger than 95%, i.e. our fixation threshold. Panels **C** and **D** of figure 2 show  $P_{fix}(f)$  as a function of  $f$  for NA and HA. For both proteins, the probability of fixation of a new mutation at frequency  $f$  is close to  $f$  itself, that is  $P_{fix}(f) \simeq f$ . This result is exactly what is expected in a population evolving in the absence of selection. A mutation or trait appearing at frequency  $f$  is shared by  $f \cdot N$  individuals, and the probability for one of them to become the ancestor of all the future population is  $f \cdot N/N = f$ . Thus, the probability of this mutation or trait to fix in the population is equal to its current frequency, a case which we will refer to as the neutral expectation. Panel **C** of figure 2 indicates that mutations in the surface proteins of A/H3N2 influenza are in good agreement with the neutral expectation, while those in A/H1N1pdm show only small deviations from it. Pierre “limited but clear” instead of “small”? In both cases, the probability of fixation seems to be mainly dictated by the current frequency  $f$  at which the mutation is observed.

This dynamics is in apparent contradiction with evidence that influenza surface proteins are under strong selective pressure to evade human immune responses (Bhatt et al., 2011). If strong selection was present, we would expect rising amino acid mutations to fix at a distinc-tively higher frequency than the one at which they are measured. In an extreme case where most trajectories would be clean sweeps,  $P_{fix}(f)$  should be close to 1 for

all but very small values of  $f$ .

Next, we searched for features of mutations that allow prediction of fixation beyond frequency by dividing frequencies into categories that deviate from the diagonal in panels **C** and **D** of figure 2. We first turn to the *Local Branching Index* (LBI), a quantity calculated for each node in a phylogenetic tree that indicates how dense the branching of the tree is around that node. LBI has previously been successfully used as a predictor of the future population of influenza (Neher et al., 2014), and was shown to be a proxy for fitness of leaves or ancestral nodes in mathematical models of evolution. Here, we define the LBI of a mutation at date  $t$  as the average LBI of strains that carry this mutation and that were sampled in the time bin corresponding to  $t$ . Panel **A** of figure 3 shows fixation probability for HA mutations with LBI in the top or bottom half of the distribution. Both groups have identical probability of fixation, suggesting that LBI carries very little information on the probability of fixation of a mutation.

Next, we focused on previously reported antigenic sites in the A/H3N2 HA protein, referred to as *epitope* positions. Mutations at these position might mediate immune escape and are therefore likely under strong selection and show sweep-like behavior. We used four lists of relevant epitope positions from different sources comprising from 7 to 129 positions in the sequence of the HA1 protein (Shih et al., 2007; Koel et al., 2013; Luksza and Lässig, 2014; Wolf et al., 2006). Panel Fig. 3**B** shows fixation probability as a function of frequency for the four lists of epitopes. Only mutations at the 7 epitope sites reported in (Koel et al., 2013) have higher chances of fixation than expected by chance. No clear difference is found for the lists by Luksza and Lässig (2014); Wolf et al. (2006), while positions from Shih et al. (2007) show lower chances of fixation. One should also note that many of these positions were determined post-hoc and might be enriched for positions that experienced rapid substitutions before the publication of the respective studies.

Two ways of categorising mutations, however, suggest some power to predict fixation. In panel Fig. 3**C**, we split trajectories into those occurring at binary positions where only two amino acid variants co-circulate and non-binary positions with more than two variants. Novel variants at non-binary positions, i.e. ones for which competition between three amino acids or more has occurred at least once, have a higher chance of fixation. In panel **D**, we separated mutations that appear more than once or only once in the reconstructed tree (see methods), and found that the former fix more often. Panels **C** and **D** show that it is possible to gain some information on the chance of fixation of a particular mutation, as was done in panel **B**. However, the predictive power remains small, with the “top” curves in panels **C&D** being very close to the diagonal.

We conduct the same analysis on A/H1N1pdm

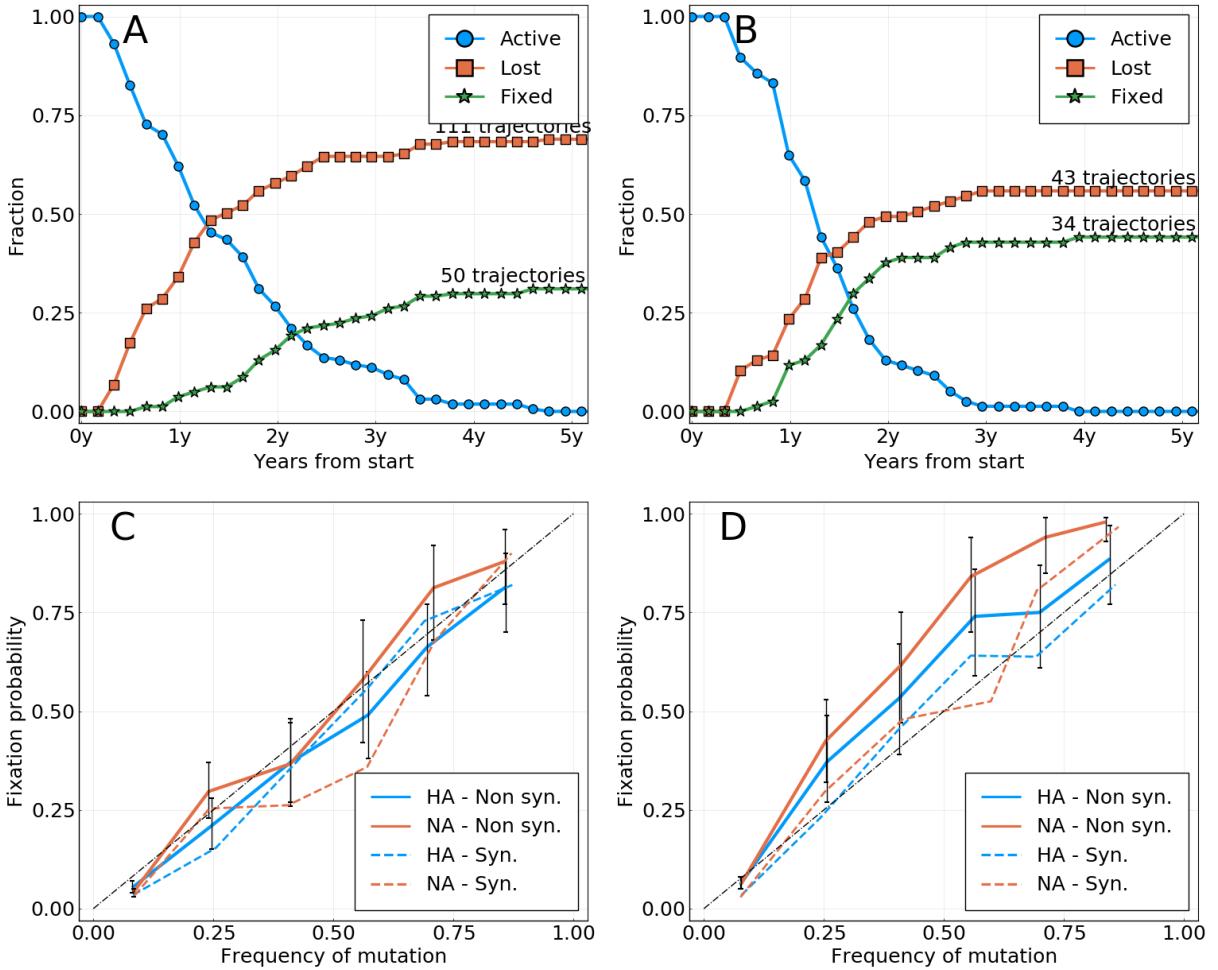


FIG. 2 **A:** Activity of all rising frequency trajectories seen above 25% frequency for A/H3N2 HA and NA. **B:** Same as **A** for A/H1N1. **C:** Probability of fixation of a mutation (amino acid or synonymous)  $P_{fix}(f)$  as a function of the frequency  $f$  at which it is measured, for A/H3N2 HA and NA. Only new mutations are considered, *i.e.* mutations that were absent in the past. The diagonal dashed line is the expectation from a neutrally evolving population. Colored dashed lines represent synonymous mutations. Colored solid lines represent amino acid mutations. Error bars represent a 95% confidence interval. **D:** Same as **C** for A/H1N1.

influenza, with results shown in figure S15. Results are qualitatively similar to those obtained for A/H3N2, with LBI giving little information and mutations at non-binary positions having a higher chance of fixation. Panel D differs between figures 3 and S15, with convergent evolution giving less information on fixation in the latter case. However, this could be due to the shorter time period over which A/H1N1pdm evolved, resulting in a shorter tree and less possibilities of convergent evolution. Indeed, error bars for mutations appearing multiple times in D of figure S15 are relatively large, indicating a lower amount of trajectories.

Since influenza is seasonal in temperate regions, geographic spread and persistence might be predictive of the success of mutations. We quantify geographic spread of a mutation by the entropy of its frequency distribution

across regions (see methods) and its persistence by the age of the trajectory by the time it reaches frequency  $f$ . Figures S16 and S17 show the fixation probabilities as a function of observed frequency for mutations classified according to these scores. The two scores also allow a quantitatively moderate distinction between mutations: for a given frequency  $f$ , mutations found in many regions or those that are older (in the sense that they have taken more time to reach frequency  $f$ ) tend to fix more often than geographically localized mutations or more recent ones, but the effect is small. These two scores are in fact correlated, with older trajectories representing mutations that are more geographically spread, as can be seen in figure S18 of SM. However, it is important to note that sampling biases and heterogeneity across time and space (see supplementary figures S9 and S10) make answering such specific hypothesis challenging. Frequency of

440 mutations might thus be amplified through different sampling biases, making the connection between geographic spread, seasonality and mutation frequency non-trivial to measure.

#### 444 Simulations of models of adaptation

445 The results shown in figures 2 and 3 are difficult to reconcile with the idea that seasonal influenza virus evolution is driven by rapid directed positive selection. One possible explanation for the weakly predictable behaviour of mutations (beyond their current frequency) might be tight genetic linkage inside each segment and strong competition between different adaptive mutations (Strelkowa and Lässig, 2012; Neher and Shraiman, 2011). We design a simple model of population evolution based on the `ffpopsim` simulation software to test this hypothesis (Zanini and Neher, 2012). The model represents a population of binary genomes of length  $L = 200$  evolving in a fitness landscape that changes through time.

446 First, we use an additive fitness function, with sequence  $(x_1 \dots x_L)$  having a fitness  $\sum_i h_i x_i$ . This implies that for a given genome position  $i$ , the trait  $x_i = 1$  is favored if  $h_i > 0$  whereas  $x_i = -1$  is favored if  $h_i < 0$ . All  $h_i$ 's have the same magnitude, and only their signs matter. Every  $\Delta t$  generations, we randomly choose a position  $i$  and flip the sign of  $h_i$ , effectively changing the fitness landscape. Individuals in the population now have the opportunity to make an adaptive mutation at site  $i$  giving them a fitness advantage  $2|h|$ . A “flip” at position  $i$  of the fitness landscape will decrease fitness of all individuals that carried the adapted variant at position  $i$  and increases the fitness of those that happened to carry a deleterious variant.

447 To increase competition between genomes, we designed a second model that includes epistasis. Once again, the baseline fitness of a genome is an additive function, this time with values of  $h_i$  that do not change through time. In addition, we added a component that mimics immune selection. Every  $\Delta t$  generation, we now introduce “antibodies” that target a specific sub-sequence of length  $l = 5$ , noted  $(x_{i_1}^{ab}, \dots, x_{i_l}^{ab})$ . The positions  $(i_1 \dots i_l)$  are chosen at random, while the targeted sub-sequence is the dominant state at each position. Genomes that include the exact sub-sequence targeted by the antibody suffer a strong fitness penalty. However, a single mutation away from that sub-sequence removes this penalty completely, resulting in a fitness landscape with very strong epistasis. This has the effect of triggering a strong competition between adaptive mutations: for a given antibody,  $l = 5$  possible mutations are now adaptive, but combinations of these mutations do not bring any fitness advantage.

448 Having simulated populations in these two fitness landscapes, we perform the same analysis of frequency trajectories as for the real influenza data. Figure S20 of the SM

449 shows the  $P_{fix}(f)$  as a function of  $f$  for the two models and for different values of the inverse rate of change  $\Delta t$  of the fitness landscape. For all models, this curve deviates significantly from the diagonal. This is most evident for the case of a simple additive fitness landscape that changes rarely  $\Delta t = 1000$ : rising mutations almost always fix in the population, with  $P_{fix}(f) \simeq 1$  for any  $f$  larger than a few percent. This is corroborated by visual inspection of the trajectories, which shows that evolution in this regime is driven by regular selective sweeps that take a typical time of  $\sim 400$  generations. In other regimes, with smaller  $\Delta t$  or with strong epistatic competition,  $P_{fix}(f)$  is reduced and closer to the diagonal. However, it takes an extremely fast changing fitness landscape to push  $P_{fix}$  close to the diagonal: with  $\Delta t = 10$ , that is about 40 changes to the fitness landscape in the time it would take a selective sweep to go from 0% to fixation,  $P_{fix}(f)$  differs from  $f$  in a way that is comparable to what is observed in A/H1N1pdm influenza.

450 These models are not meant to be accurate models of influenza viruses evolution. But figure S20 does show that the patterns observed in influenza virus evolution are only reproduced by models of adapting populations when pushing clonal competition to extreme values. We conclude that the pattern in figure 2 may not be a straightforward manifestation of genetic linkage and clonal interference, but that some more intricate interplay of epidemiology, seasonality, human immunity and chance gives rise to the weakly predictable yet strongly selected evolutionary dynamics of IAVs.

#### Why do predictions work?

451 The statistics of frequency trajectories seem to be in conflict with the notion that influenza evolution is predictable. Likewise, the LBI, a quantity that correlates with fitness in mathematical models and is used to predict future influenza populations (Neher et al., 2014), does not seem to contain any information on whether a specific mutation is going to fix or not, see figure 3. To resolve this conundrum, we first note that the criterion by which predictive power for influenza was measured in (Neher et al., 2014) was the distance between the strain with the highest LBI and the future population, not the ability of the LBI to predict dynamics. The distance was compared to the average distance between the present and future population, as well as the post-hoc optimal representative and the future.

452 To quantify the ability of the LBI and other measures to pick good representatives of the future, we construct a large tree of HA sequences with 100 sequences in non-overlapping time bins of 4 months from year 2003 to 2019 (a total of 4402 as some 4 month intervals contain less than 100 sequences). Each time bin is considered as a snapshot of the A/H3N2 influenza population and we will

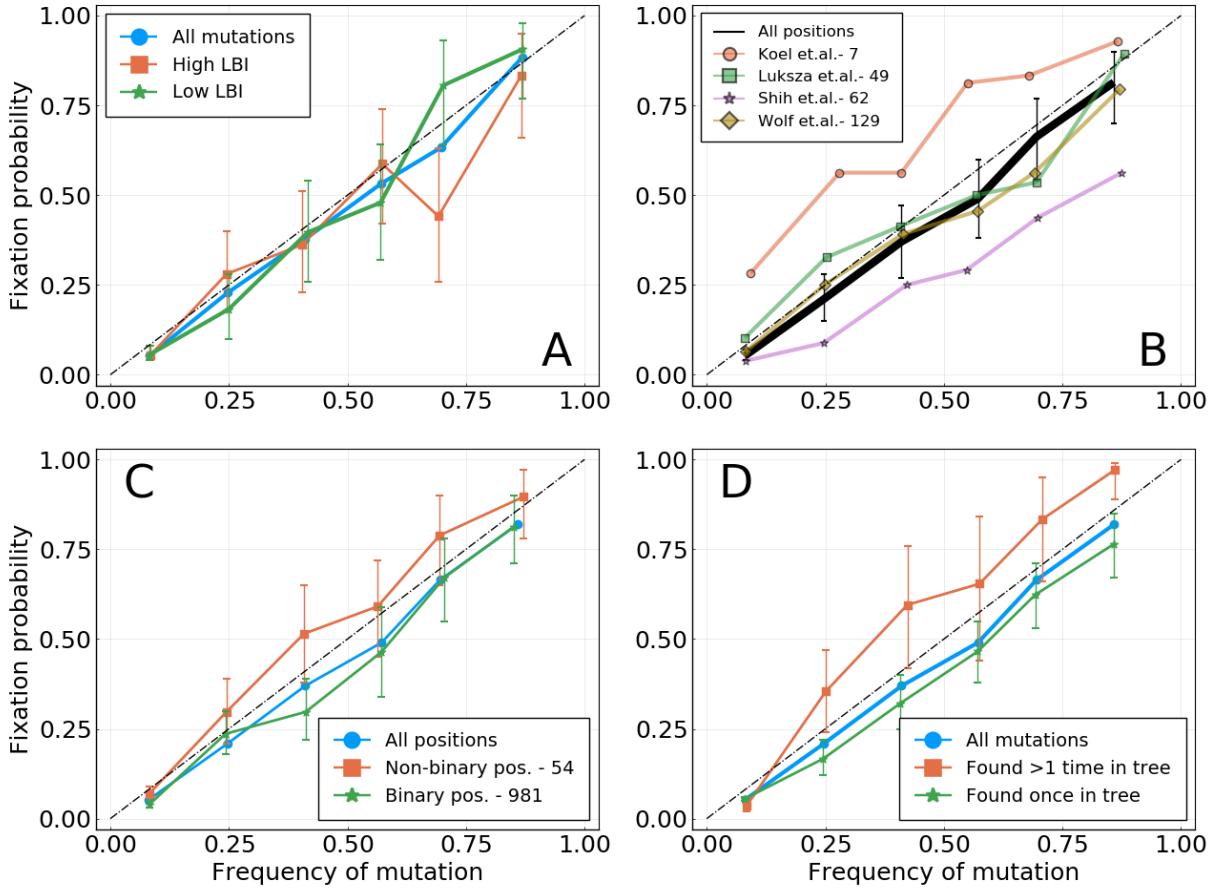


FIG. 3 Fixation probability  $P_{fix}(f)$  as a function of frequency, for A/H3N2 influenza. Figure S15 shows the same analysis for A/H1N1. **A:** HA mutations with higher or lower LBI values, based on their position with respect to the median LBI value. **B:** Different lists of epitope positions in the HA protein. The authors and the number of positions is indicated in the legend. **C:** HA and NA mutations for binary positions, *i.e.* positions for which we never see more than two amino acids in the same time bin. **D:** HA and NA mutations that appear once or more than once in the tree for a given time bin.

refer to sequences in time bin  $t$  as the population of the present. From this present population, we predict future populations in time bin  $t + \Delta t$ , using only sequences in time bin  $t$  and before.

To assess the ability of the LBI to pick a close representative of the future, we compute the LBI of each node of one time bin in the tree using only the leaves that belong to that time bin. The top panel in figure 4 shows the hamming distance of the strain with the highest LBI to future populations at different  $\Delta t$  along with the same distance for a randomly chosen strain. The figure shows the distance averaged over all possible values of  $t$  for  $\Delta t$  between 0 and 32 months, giving us an average efficiency of a predictor over 16 years of influenza evolution.

The strain with the highest LBI is consistently closer to the future than the average strain by about 1-2 amino acids, while the overall distance increases linearly due to the continuous evolution of the population. We hence reproduce previous results showing that the LBI picks closer than average representatives (Neher et al., 2014).

To investigate whether this apparent success is due to the ability of the LBI to predict fitness or not, we explored a different predictor: the amino acid consensus sequence of the present population (see Methods for a definition of the consensus sequence). The choice is motivated by the fact that it can be shown to be the best possible long term predictor for a neutrally evolving population in terms of Hamming distance (see SM section .1). Figure 4 shows that the consensus sequence is in fact a equally good or even slightly better representative of the future than the sequence with highest LBI (note that the consensus sequence does *not* necessarily exist in the population).

This near equivalence of the consensus and the strain with highest LBI can be explained as follows: The LBI tends to be high for nodes in a tree that are close to the root of a dense and large clade. A typical sample of influenza HA sequences fall into a small number of recognizable clades, and the strains with maximal LBI will often be close to the root of the largest of those clades. This root of the largest clade will often be close to the

586 consensus of the whole population, explaining the similar<sup>639</sup>  
 587 distance patterns. To test that hypothesis, we measure the<sup>640</sup>  
 588 hamming distance from the sequence of the top LBI strain<sup>641</sup>  
 589 to the consensus sequence for populations of all time bins<sup>642</sup>  
 590 Panel **B** of figure 4 shows these distances, scaled with<sup>643</sup>  
 591 respect to an average strain (details in caption). It clearly<sup>644</sup>  
 592 shows that the top-LBI strain and the consensus sequence<sup>645</sup>  
 593 are indeed quite similar: out of 48 time bins, only once<sup>646</sup>  
 594 is the sequence of the top-LBI strain farther away from<sup>647</sup>  
 595 the consensus than the average sequence is. Moreover<sup>648</sup>  
 596 the sequence of the top-LBI strain *exactly* matches the<sup>649</sup>  
 597 consensus in 19 cases.

650  
 651 while others gave limited information on fixation. Despite  
 652 the lack of predictability of mutation frequency trajectories,  
 653 influenza surface proteins show strong signatures of  
 654 selection (Bhatt et al., 2011; Strelkowa and Lässig, 2012).

655 Methods for predicting the future evolution of influenza  
 656 either construct explicit fitness models (Luksza and Lässig,  
 657 2014; ?), use historical patterns of evolution (Luksza and  
 658 Lässig, 2014; Bush et al., 1999), phenotypic assays (Neher  
 659 et al., 2016; Steinbrück and McHardy, 2012), or dynamic  
 660 or phylogenetic patterns (Neher et al., 2014; Klingen et al.,  
 661 2018b). The goal of these methods is to pick strains that  
 662 are good representatives of future populations and could  
 663 serve as vaccine candidates (Morris et al., 2018).

664 The low power to predict frequency dynamics or fixation  
 665 naturally triggers the question why the above methods  
 666 have been found to work. Picking representatives of the  
 667 future and predicting frequency dynamics are distinct  
 668 objectives and success at the former (as compared to ran-  
 669 dom picks) is not necessarily inconsistent with a lack of  
 670 predictable dynamics. In fact, (?) reports that the rate  
 671 at which the frequency of a strain changes is often a poor  
 672 predictor – consistent with our observations here. But  
 673 despite the fact that future frequencies are not predicted  
 674 by the LBI, the strain with the highest LBI in the pop-  
 675 ulation is a better predictor of the future population than  
 676 a randomly picked one. While the LBI was shown to be  
 677 a correlate of relative fitness and be predictive of fixation  
 678 in mathematical models of evolution (Neher et al., 2014),  
 679 it does not seem to be predictive of influenza  
 680 evolution because it measures fitness from genealogical  
 681 structure. Instead, we believe it picks closer than average  
 682 strains simply because it has the tendency to be maximal  
 683 at the base of large and dense clades. These basal geno-  
 684 types are closer to the future populations than the current  
 685 tips of the tree and hence a better predictor on average.  
 686 The consensus sequence of all present strains performs  
 687 slightly but consistently better than picking the strain  
 688 with the highest LBI. The consensus sequence is the best  
 689 possible predictor for a neutrally evolving population, and  
 690 does not attempt to model fitness in any way.

691 At the same time, influenza virus phylogenies show clear  
 692 deviations from those expected from the neutral Kingman  
 693 coalescent, similar to those expected under Bolthausen-  
 694 Sznitman coalescent (BSC) processes that are generated  
 695 by traveling wave models of rapid evolution (Neher and  
 696 Hallatschek, 2013; Desai et al., 2013). The correspon-  
 697 dence between the BSC and traveling wave models comes  
 698 from transient exponential amplification of fit strains be-  
 699 fore these fitness differences are wiped out by further  
 700 mutation. This exponential amplification generates long-  
 701 tailed effective offspring distributions which in turn can  
 702 lead to genealogies described by the BSC (Neher and  
 703 Hallatschek, 2013; Schweinsberg, 2003). Many processes  
 704 other than selection, including seasonality and spatio-  
 705 temporal heterogeneity, can generate effective long tailed  
 706 offspring distributions even in absence of bona-fide fitness

## 598 DISCUSSION

599 Predicting the trajectory of a mutation requires (i)<sup>655</sup>  
 600 significant fitness difference between genomes carrying<sup>656</sup>  
 601 different variants at the site and (ii) a selection pressure<sup>657</sup>  
 602 that changes slowly over time. Under such conditions, it is<sup>658</sup>  
 603 expected that frequency trajectories will show a persistent<sup>659</sup>  
 604 behavior which would make them predictable for some<sup>660</sup>  
 605 time. However, we could find only limited evidence for<sup>661</sup>  
 606 such persistent behavior in the past 19 years of IAV<sup>662</sup>  
 607 evolution. This lead us to conclude that (i) influenza<sup>663</sup>  
 608 virus evolution is qualitatively different from models of<sup>664</sup>  
 609 rapidly adapting population (despite clear evidence for<sup>665</sup>  
 610 frequent positive selection), and (ii) previous methods to<sup>666</sup>  
 611 predict influenza evolution work primarily because they<sup>667</sup>  
 612 pick strains that represent the future well, not because<sup>668</sup>  
 613 they predict future dynamics.

614 The primary focus in this work was the investigation<sup>670</sup>  
 615 of frequency trajectories of new amino acid mutations. In<sup>671</sup>  
 616 the short term, we found that on average the direction of<sup>672</sup>  
 617 trajectory does not persist for longer than a few months<sup>673</sup>  
 618 Indeed, the average trajectory in figure 1 takes a sharp<sup>674</sup>  
 619 turn when going from  $t < 0$  to  $t > 0$ , instead of showing<sup>675</sup>  
 620 “inertia”. This suggests that selective sweeps are not<sup>676</sup>  
 621 representative of typical trajectories. **Pierre: We could**<sup>677</sup>  
 622 **cite the Illingworth paper here. They do show that as**<sup>678</sup>  
 623 **well, in a different way. The fact many trajectories are not**<sup>679</sup>  
 624 **sweeps is expected when there is strong linkage. What**<sup>680</sup>  
 625 **cannot be easily explained by linkage only is the complete**<sup>681</sup>  
 626 **lack of inertia on average (at least I think?), and the**<sup>682</sup>  
 627  **$P_{fix}(f) = f$ .**

628 On a longer timescale, we investigated the probability<sup>684</sup>  
 629 that a novel mutation observed at frequency  $f$  fixes. In<sup>685</sup>  
 630 neutral models of evolution this probability equals  $f$ , while<sup>686</sup>  
 631 it should be higher or lower than  $f$  for mutations with<sup>687</sup>  
 632 a beneficial or deleterious effect on fitness, respectively<sup>688</sup>  
 633 However, in the case of influenza, this probability differs<sup>689</sup>  
 634 little from  $f$ , making current frequency the best predictor<sup>690</sup>  
 635 for fixation. In figure 3, we split trajectories into groups<sup>691</sup>  
 636 for which we expected  $P_{fix}$  to deviate from  $f$ . Many of<sup>692</sup>  
 637 these splits, such as high/low LBI or epitope/non-epitope<sup>693</sup>  
 638 positions, did not result in an increased predictability<sup>694</sup>

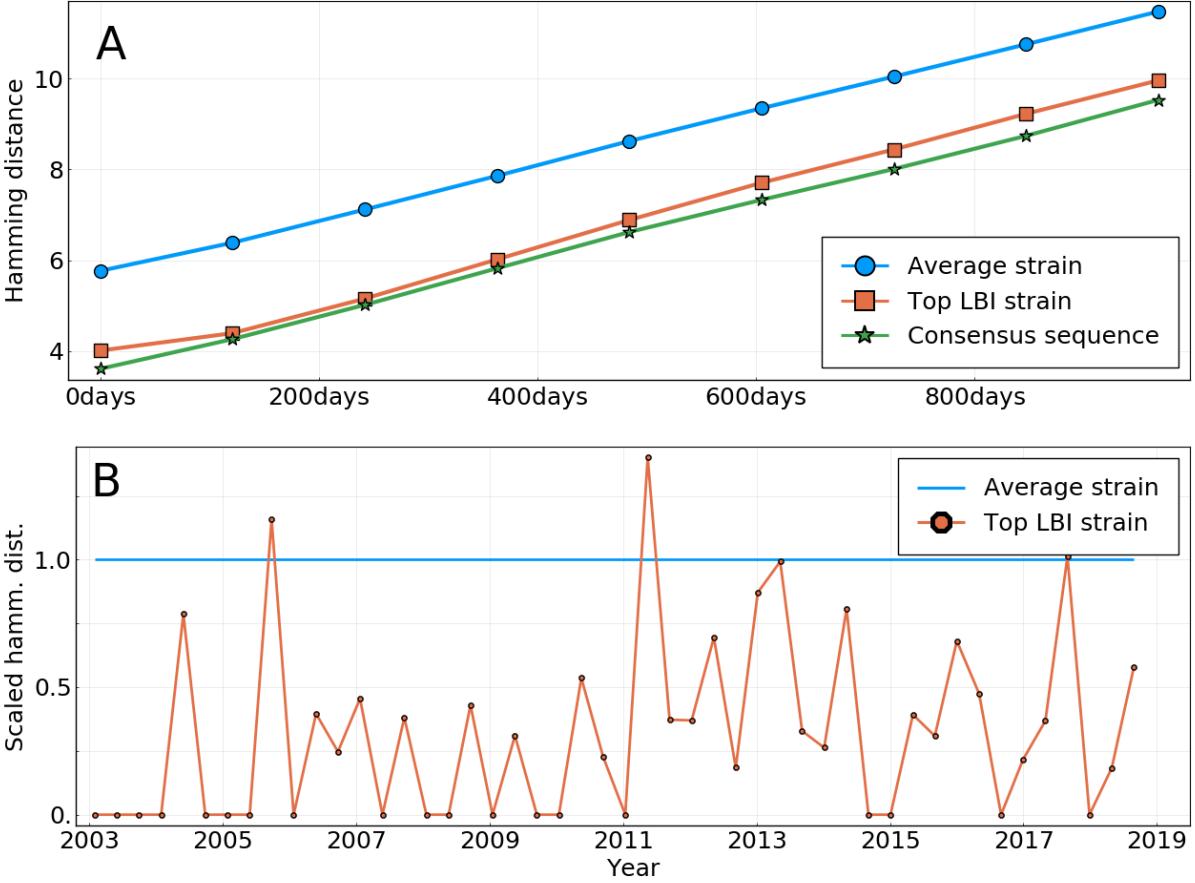


FIG. 4 **A:** Average Hamming distance of the sequences of different predictors to HA sequences of future influenza populations, themselves averaged over all “present” populations from year 2003 to 2019. Predictors are: a randomly picked sequence in the present population; the sequence of the strain with the highest LBI in the present population; the consensus sequence of the present population. **B:** Scaled Hamming distance between the sequence of the top LBI strain and the consensus sequence for populations at different dates. The scaling is such that for each date, the Hamming distance between a strain from the population and the consensus is on average 1. The strain with the highest LBI is almost always closer to the consensus sequence than the average strain.

695 differences, which might explain ladder-like non-Kingman<sup>712</sup> phylogenetic trees.

## METHODS

### 713 Data and code availability

The sequences used are obtained from the GISAID database (Shu and McCauley, 2017). Strain names and accession numbers are given as tables in two supplementary files. Outliers strains listed at <https://github.com/PierreBarrat/FluPredictability/src/config> were removed.

The code used to generate the figures presented here is available at <https://github.com/PierreBarrat/FluPredictability>.

### Frequency trajectories

For a set of sequences in a given time bin, we compute frequencies of amino acids at each position by simple

697 A recent preprint proposed that influenza virus evolution<sup>714</sup> is primarily limited by an asynchrony between<sup>715</sup> population level selection and generation of new variants<sup>716</sup> within infected hosts (Morris et al., 2020). Along these<sup>717</sup> lines, it is possible that the A/H3N2 population readily<sup>718</sup> responds once population level selection is high enough by<sup>719</sup> giving rise to essentially equivalent variants. Furthermore,<sup>720</sup> selection might cause the rapid rise of a novel variant to<sup>721</sup> macroscopic frequencies (observable in a global sample)<sup>722</sup> but its benefit rapidly “expires” because competing vari-<sup>723</sup> ants catch up and/or it mediates immune escape only<sup>724</sup> to a small fraction of the population. These consider-<sup>725</sup> ations might explain the disconnect between models of<sup>726</sup> rapid adaptation and the frequency dynamics observed<sup>727</sup> in influenza virus populations.

726 counting. We make the choice of not applying any smoothing method in an attempt to be as close to the data and “model-less” as possible. This is especially important for the short term prediction of frequency trajectories, as estimations of the “persistence time” of a trajectory might be biased by a smoothing method.

727 We compute frequency trajectories based on the frequencies of amino acids. A trajectory begins at time  $t$  if an amino acid is seen under the lower frequency threshold of 5% (resp. above the higher threshold of 95%) for the two time bins preceding  $t$ , and above this lower threshold (resp. below the higher threshold) for time bin  $t$ . It ends in the reciprocal situation, that is when the frequency is measured below the lower threshold (resp. above the higher threshold) for two time bins in a row.

728 In order to avoid estimates of frequencies that are too noisy, we only keep trajectories that are based on a population of at least 10 sequences for each time bin. As said in the Results section, we also restrict the analysis to trajectories that begin at a 0 frequency, in part to avoid double counting. We find a total of 460 such trajectories. However, only 106 reach a frequency of 20%, on which figure 2 is based for instance.

729 Note that the fact that we use samples of relatively small sizes – at least for some time bins – leads to biases in the estimation of frequencies. We show in Supplementary Material that these biases are generally small and do not induce any qualitative changes to results presented here.

## 730 Local Branching Index

731 LBI was introduced in (Neher et al., 2014) as an approximation of fitness in populations evolving under persistent selective pressure that is fully based on a phylogenetic tree. It relies on the intuition that the tree below high-fitness individuals will show dense branching events, whereas absence of branching is a sign of low-fitness individuals. Quantitatively, the LBI  $\lambda_i(\tau)$  of a node  $i$  is the integral of all of the tree’s branch length around  $i$ , with an exponentially decreasing weight  $e^{-t/\tau}$  with  $t$  being the branch length. When considering a time binned population, the LBI is computed once for each time bin by considering only the leaves of the tree that belong to the time bin. This means that only branches that ultimately lead to a leaf that belongs to the time bin are considered in the integration.

732  $\tau$  is the time scale for which the tree is informative of the fitness of a particular node. Here, we use a value of  $\tau$  equal to a tenth of  $T_C \simeq 6$  years, the coalescence time for influenza A/H3N2 strains, converted to units of tree branch length through the average nucleotide substitution rate ( $\simeq 4 \cdot 10^{-3}$  substitutions per site per year for HA). We have observed that given our method to predict the future from present populations corresponding to time bins of 4 months, changing the value of  $\tau$  has little effect.

733 on the pick of the top LBI strain. By retrospectively optimizing its value, it is possible to reduce the average distance to the population 2 years ahead by  $\sim 0.25$  amino acids on average, making the LBI method almost as good as the consensus on figure 4.

## 734 Measuring the geographical spread of a mutation

735 For a mutation  $X$  we define its regional distribution using the numbers  $n_r(X)$  that represent the number of sequences sampled in region  $r$  that carry  $X$ . Regional weights are then defined as

$$w_r(X) = \frac{n_r(X)}{\sum_r n_r(X)}.$$

736 We can then measure the geographical spread  $G(X)$  of  $X$  by using the Shannon entropy of the probability distribution  $w_r(X)$ :

$$G(X) = \sum_r w_r(X) \log(w_r(X)).$$

737  $G(X)$  is a positive quantity that is larger when  $X$  is equally present in many regions, and equal to zero when  $X$  is concentrated in only one region.

738 Region used are the ones defined in the Nextstrain tool (Hadfield et al., 2018). Those are North America, South America, Europe, China, Oceania, Southeast Asia, Japan & Korea, South Asia, West Asia, and Africa.

## 739 Assigning a fitness to trajectories

### 740 Consensus sequence

741 Given a set of  $N$  sequences  $(\sigma^1, \dots, \sigma^N)$  based on an alphabet  $\mathcal{A}$  (e.g.  $\mathcal{A}$  has 20 elements for amino acids, 4 for nucleotides), we can define a profile distribution  $p_i(a)$  by the following expression:

$$p_i(a) = \sum_{n=1}^N \delta_{\sigma_i^n, a}$$

742 where  $i$  is a position in the sequence,  $\sigma_i^n$  the character appearing at position  $i$  in sequence  $\sigma^n$ ,  $a$  a character of the alphabet and  $\delta$  the Kronecker delta. The profile  $p_i(a)$  simply represents the fraction of sequences which have character  $a$  at position  $i$ .

743 We then simply define the consensus sequence  $\sigma^{cons}$  such that

$$\sigma_i^{cons} = \operatorname{argmax}_a p_i(a).$$

744 In other words, the consensus sequence is the one that has the dominant character of the initial set of sequences at each position.

797 Earth Mover's Distance 82

In order to measure the distance of several predictor sequences to the future population, we rely on the *Earth Mover's Distance* (EMD), a metric commonly applied in machine learning to compare collections of pixels or words (Rubner et al., 1998; Kusner et al., 2015). Here, we apply it to compute the distance between the sequences of two populations, noted as  $\mathcal{X} = \{(x^n, p^n)\}$  and  $\mathcal{Y} = \{(y^m, q^m)\}$  with  $n \in \{1 \dots N\}$  and  $m \in \{1 \dots M\}$ . In this notation,  $x^n$  and  $y^m$  are sequences, and  $p^n$  and  $q^m$  are the frequencies at which these sequences are found in their respective populations. For convenience, we also define  $d_{mn} = H(x^n, y^m)$  as the Hamming distance between pairs of sequences in the two populations.

We now introduce the following functional

$$F(\mathbf{w}) = \sum_{n,m} d_{nm} w_{nm},$$

with  $\mathbf{w} = \{w_{nm}\}$  being a matrix of positive weights. The EMD between the two populations  $\mathcal{X}$  and  $\mathcal{Y}$  is now defined as the minimum value of function  $F$  under the conditions

$$\sum_{n=1}^N w_{nm} = q^m, \quad \sum_{m=1}^M w_{nm} = p^n, \text{ and } w_{nm} \geq 0$$

Intuitively, the weight  $w_{nm}$  tells us how much of sequence  $x^n$  is “moved” to sequence  $y^m$ . The functional  $F$  sums all of these moves and attributes them a cost equal to the Hamming distance  $d_{nm}$ . The conditions on weights in  $\mathbf{w}$  ensure that all the weight  $p^n$  of  $x^n$  is “moved” to elements in  $\mathcal{Y}$  and vice versa.

The minimization is easily performed by standard linear optimization libraries. Here, we use the Julia library JuMP (Dunning et al., 2017).

## I. FUNDING

This work was funded in part by NIAID (R01AI127893-<sup>868</sup>  
01) and the SNF (310030-188547). <sup>869</sup>

## REFERENCES

- evolution. *Proceedings of the National Academy of Sciences*, 104(15):6283–6288, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0701396104. URL <https://www.pnas.org/content/104/15/6283>.

Samir Bhatt, Edward C. Holmes, and Oliver G. Pybus. The Genomic Rate of Molecular Adaptation of the Human Influenza A Virus. *Molecular Biology and Evolution*, 28(9):2443–2451, September 2011. ISSN 0737-4038. doi:10.1093/molbev/msr044. URL <https://academic.oup.com/mbe/article/28/9/2443/1007907>.

Björn F. Koel, David F. Burke, Theo M. Bestebroer, Stefan van der Vliet, Gerben C. M. Zondag, Gaby Vervaet, Eugene Skepner, Nicola S. Lewis, Monique I. J. Spronken, Colin A. Russell, Mikhail Y. Eropkin, Aeron C. Hurt, Ian G. Barr, Jan C. de Jong, Guus F. Rimmelzwaan, Albert D. M. E. Osterhaus, Ron A. M. Fouchier, and Derek J. Smith. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science*, 342(6161):976–979, 2013. ISSN 0036-8075. doi:10.1126/science.1244730. URL <https://science.sciencemag.org/content/342/6161/976>.

Dylan H Morris, Katelyn M Gostic, Simone Pompei, Trevor Bedford, Marta Luksza, Richard A Neher, Bryan T Grenfell, Michael Lässig, and John W McCauley. Predictive modeling of influenza shows the promise of applied evolutionary biology. *Trends in microbiology*, 26(2):102–118, 2018.

Thorsten R. Klingen, Susanne Reimering, Carlos A. Guzmán, and Alice C. McHardy. In Silico Vaccine Strain Prediction for Human Influenza Viruses. *Trends in Microbiology*, 26(2):119–131, February 2018a. ISSN 0966-842X. doi:10.1016/j.tim.2017.09.001. URL <http://www.sciencedirect.com/science/article/pii/S0966842X17302068>.

Peter Bogner, Ilaria Capua, David J Lipman, and Nancy J Cox. A global initiative on sharing avian flu data. *Nature*, 442(7106):981–981, 2006.

Yuelong Shu and John McCauley. Gisaid: Global initiative on sharing all influenza data—from vision to reality. *Euro-surveillance*, 22(13), 2017.

Andrew Rambaut, Oliver G. Pybus, Martha I. Nelson, Cécile Viboud, Jeffery K. Taubenberger, and Edward C. Holmes. The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453(7195):615–619, May 2008. ISSN 1476-4687. doi:10.1038/nature06945. URL <https://www.nature.com/articles/nature06945>.

Marta Luksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507(7490):57–61, March 2014. ISSN 1476-4687. doi:10.1038/nature13087. URL <https://www.nature.com/articles/nature13087>. Number: 7490 Publisher: Nature Publishing Group.

L. Steinbrück, T. R. Klingen, and A. C. McHardy. Computational prediction of vaccine strains for human influenza a (h3n2) viruses. *Journal of Virology*, 88(20):12123–12132, 2014. ISSN 0022-538X. doi:10.1128/JVI.01861-14. URL <https://jvi.asm.org/content/88/20/12123>.

Richard A. Neher, Trevor Bedford, Rodney S. Daniels, Colin A. Russell, and Boris I. Shraiman. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 113(12):E1701–1709, March 2016. ISSN 1091-6490 0027-8424. doi: 10.1073/pnas.1525578113.

# Richard A Neher, Colin A Russell, and Boris I Shraiman. Predicting evolution from the shape of genealogical trees.

- 886 *eLife*, 3, November 2014. ISSN 2050-084X. doi:10.1371/journal.pcbi.1002492.  
 887 10.7554/eLife.03568. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4227306/>.  
 888
- 889 Natalja Strelkowa and Michael Lässig. Clonal Interference in the Evolution of Influenza. *Genetics*, 192(2):671–682, October 2012. ISSN 0016-6731, 1943-2631. doi:10.1534/genetics.112.143396. URL <http://www.genetics.org/content/192/2/671>.  
 890
- 891 Yuri I Wolf, Cecile Viboud, Edward C Holmes, Eugene V Koonin, and David J Lipman. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biology Direct*, 1:34, October 2006. ISSN 1745-6150. doi:10.1186/1745-6150-1-34. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1647279/>.  
 892
- 893 Juhye M Lee, Rachel Eguia, Seth J Zost, Saket Choudhary, Patrick C Wilson, Trevor Bedford, Terry Stevens-Ayers, Michael Boeckh, Aeron C Hurt, Seema S Lakdawala, Scott E Hensley, and Jesse D Bloom. Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. *eLife*, 8:e49324, August 2019. ISSN 2050-084X. doi:10.7554/eLife.49324. URL <https://doi.org/10.7554/eLife.49324>. Publisher: eLife Sciences Publications, Ltd.  
 894
- 895 Motoo Kimura. Diffusion Models in Population Genetics. *Journal of Applied Probability*, 1(2):177–232, 1964. ISSN 0021-9002. doi:10.2307/3211856. URL <http://www.jstor.org/stable/3211856>.  
 896
- 897 Le Yan, Richard A Neher, and Boris I Shraiman. Phylodynamic theory of persistence, extinction and speciation of rapidly adapting pathogens. *eLife*, 8:e44205, September 2019. ISSN 2050-084X. doi:10.7554/eLife.44205. URL <https://doi.org/10.7554/eLife.44205>. Publisher: eLife Sciences Publications, Ltd.  
 898
- 899 R. A. Neher and B. I. Shraiman. Genetic draft and quasi neutrality in large facultatively sexual populations. *Genetics*, 188(4):975–996, August 2011. ISSN 1943-2631 0016-6731. doi:10.1534/genetics.111.128876.  
 900
- 901 Fabio Zanini and Richard A. Neher. FFPopSim: an efficient forward simulation package for the evolution of large populations. *Bioinformatics*, 28(24):3332–3333, 10 2012. ISSN 1367-4803. doi:10.1093/bioinformatics/bts633. URL <https://doi.org/10.1093/bioinformatics/bts633>.  
 902
- 903 R. M. Bush, C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch. Predicting the evolution of human influenza A. *Science (New York, N.Y.)*, 286(5446):1921–1925, December 1999. ISSN 0036-8075.  
 904
- 905 Katherine S. Xue, Louise H. Moncla, Trevor Bedford, and Jesse D. Bloom. Within-Host Evolution of Human Influenza Virus. *Trends in Microbiology*. ISSN 0966-842X. doi:10.1016/j.tim.2018.02.007. URL <https://www.sciencedirect.com/science/article/pii/S0966842X1830043X>.  
 906
- 907 Dylan H. Morris, Velislava Petrova, Fernando W. Rossine, Edyth Parker, Bryan Grenfell, Richard Neher, Simon Levin, and Colin Russell. Asynchrony between virus diversity and antibody selection limits influenza virus evolution. preprint, Open Science Framework, April 2020. URL <https://osf.io/847p2>.  
 908
- 909 Lars Steinbrück and Alice Carolyn McHardy. Inference of Genotype–Phenotype Relationships in the Antigenic Evolution of Human Influenza A (H3N2) Viruses. *PLOS Computational Biology*, 8(4):e1002492, April 2012. ISSN 1553-7358. doi:10.1371/journal.pcbi.1002492. URL <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002492>.  
 910 Thorsten R. Klingen, Susanne Reimering, Jens Loers, Kyra Mooren, Frank Klawonn, Thomas Krey, Gülsah Gabriel, and Alice C. McHardy. Sweep Dynamics (SD) plots: Computational identification of selective sweeps to monitor the adaptation of influenza A viruses. *Scientific Reports*, 8(1):373, January 2018b. ISSN 2045-2322. doi:10.1038/s41598-017-18791-z. URL <https://www.nature.com/articles/s41598-017-18791-z>.  
 911
- 912 Richard A. Neher and Oskar Hallatschek. Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences of the United States of America*, 110(2):437–442, January 2013. ISSN 1091-6490 0027-8424. doi:10.1073/pnas.1213113110.  
 913
- 914 Michael M. Desai, Aleksandra M. Walczak, and Daniel S. Fisher. Genetic Diversity and the Structure of Genealogies in Rapidly Adapting Populations. *Genetics*, 193(2):565–585, February 2013. ISSN 0016-6731, 1943-2631. doi:10.1534/genetics.112.147157. URL <http://www.genetics.org/content/193/2/565>.  
 915
- 916 Jason Schweinsberg. Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic Processes and their Applications*, 106(1):107–139, July 2003. ISSN 0304-4149. doi:10.1016/S0304-4149(03)00028-0. URL <http://www.sciencedirect.com/science/article/pii/S0304414903000280>.  
 917
- 918 James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagunenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 05 2018. ISSN 1367-4803. doi:10.1093/bioinformatics/bty407. URL <https://doi.org/10.1093/bioinformatics/bty407>.  
 919
- 920 Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 59–66, Jan 1998. doi:10.1109/ICCV.1998.710701.  
 921
- 922 Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 957–966. JMLR.org, 2015.  
 923
- 924 Iain Dunning, Joey Huchette, and Miles Lubin. Jump: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017. doi:10.1137/15M1020575.  
 925
- 926 Julia Sigwart. Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.—Jotun Hein, Mikkel H. Schierup, and Carsten Wiuf. 2004. Oxford University Press, Oxford. xiii + 276 pp. ISBN 0-19-852996-1, £29.95 (paperback); ISBN 0-19-852995-3, £65.00 (hardback). *Systematic Biology*, 54(6):986–987, 12 2005. ISSN 1063-5157. doi:10.1080/10635150500354860. URL <https://doi.org/10.1080/10635150500354860>.  
 927

1004 **SUPPLEMENTARY MATERIAL**1005 **1. Consensus sequence as a predictor for neutrally evolving populations**

1006 We consider the case of a neutrally evolving and structure-less population, such as the one in the Wright-Fisher  
 1007 model of evolution (Sigwart, 2005). At an initial time  $t = 0$ , the population consists of  $N$  individuals with genomes  
 1008 ( $\sigma^1 \dots \sigma^N$ ) of length  $L$  (not necessarily distinct).

1009 We make two hypotheses about this population. We first suppose that *no* mutations occur during the evolution of this  
 1010 population. This may seem surprising and is of course not true in the case of influenza. This assumption is however in  
 1011 line with the fact that the object of this work is to predict the outcome of *already existing* mutations in the influenza  
 1012 population. The prediction of mutations that we have not yet seen is not in its scope. Thus, assuming that no new  
 1013 mutations take place can be seen as a simple way to model the fact that we have no information about such events.  
 1014 The second assumption is that the population evolves in a completely neutral way, meaning that the average number  
 1015 of descendants of each genome  $\sigma^n$  is the same. Let us now consider the population after it has evolved for a long time  
 1016  $t \gg T$  where  $T$  is the typical coalescence time (for the Wright-Fisher model,  $T = 2N$ ). At this point, all individuals in  
 1017 the future population will descend from a unique individual  $n_0$  in the  $t = 0$  population. Our two hypotheses now allow  
 1018 us to make two statements. First, since no new mutations are allowed, the population at  $t \gg T$  will be clonal, with all  
 1019 individuals having genome  $\sigma^{n_0}$ . Second, since the evolution is neutral and does not favour any genome in particular,  
 1020 the probability that  $\sigma^{n_0}$  is equal to a given genome  $\sigma$  is  $1/N$ . In other words, the probability that a genome at  $t = 0$   
 1021 ultimately becomes the ancestor of all the future population is equal to its frequency in the  $t = 0$  population.

1022

1023 We now try to find the genome  $\sigma$  that best predicts the future population on the long run, that is for  $t \gg T$ . Here,  
 1024 we take best to mean that the predictor minimizes  $H(\sigma, \sigma^{n_0})$  where  $H$  is the Hamming distance defined by

$$H(\sigma^a, \sigma^b) = \sum_{i=1}^L (1 - \delta_{\sigma_i^a, \sigma_i^b}), \quad (1)$$

1025 with  $\sigma_i$  being the character appearing at position  $i$  of genome  $\sigma$  and  $\delta$  the Kronecker delta. Since we do not know  $n_0$ ,  
 1026 we have to average over all its possible values.  $\sigma$  must thus minimize the following quantity:

$$\begin{aligned} \langle H(\sigma, \sigma^{n_0}) \rangle_{n_0} &= \sum_{n=1}^N H(\sigma, \sigma^n) \\ &= \sum_{i=1}^L \sum_{n=1}^N (1 - \delta_{\sigma_i, \sigma_i^n}) \end{aligned} \quad (2)$$

by using the definition of the Hamming distance. We now assume that characters at each positions of the genomes can  
 be indexed by an integer  $a$  running from 1 to  $q$ . For instance, if these were amino acid sequences, we could index the  
 20 amino acids by  $a$  running from 1 to  $q = 20$ . We rewrite the Kronecker delta in the previous expression using this  
 indexation:

$$\delta_{\sigma_i, \sigma_i^n} = \sum_{a=1}^q \delta_{\sigma_i, a} \delta_{\sigma_i^n, a}.$$

1027 We also introduce the *profile* frequencies  $p_i(a)$  of the population at time  $t = 0$ :

$$p_i(a) = \sum_{n=1}^N \delta_{\sigma_i^n, a}. \quad (3)$$

1028  $p_i(a)$  represents the frequency at which character  $a$  appears at position  $i$  in genomes of the initial population.

1029 Equation 2 now becomes

$$\begin{aligned}
 \langle H(\sigma, \sigma^{n_0}) \rangle_{n_0} &= \sum_{i=1}^L \sum_{n=1}^N \left( 1 - \sum_{a=1}^q \delta_{\sigma_i, a} \delta_{\sigma_i^n, a} \right) \\
 &= \sum_{i=1}^L \left( 1 - \sum_{a=1}^q \delta_{\sigma_i, a} p_i(a) \right) \\
 &= \sum_{i=1}^L (1 - p_i(\sigma_i))
 \end{aligned} \tag{4}$$

1030 This means that the genome  $\sigma = (\sigma_1 \dots \sigma_L)$  which best predicts the future population according to our definition is  
 1031 the one that minimizes the quantity  $(1 - p_i(\sigma_i))$  for all positions  $i$ . This obviously implies that each  $\sigma_i$  must be chosen  
 1032 as to maximize  $p_i(a)$ , that is  $\sigma_i$  must be the character that appears the most frequently at position  $i$ . Thus,  $\sigma$  must be  
 1033 the *consensus* sequence of the initial population.

## 1034 2. Predictor based on the local LBI maxima

In figure 19, we use several sequences as a predictor of the future population. Distance between two sets of sequences, *i.e.* the predictor sequences and the ones of the future population, is defined as the Earth Mover's Distance (EMD). Here, we show that for a population evolving under the same hypotheses as in section .1, the best *multiple* sequence long term predictor is again the consensus sequence with weight 1.

Let the predictor be a set of weighted sequences  $\{(s^\alpha, q_\alpha)\}$ . We again use the fact that in the long term, a unique sequence  $\sigma^{n_0}$  from the present will be the ancestor of the entire population. We want to compute the EMD from the predictor to  $\sigma^{n_0}$ , that is the EMD between the sets  $\mathcal{X} = \{(s^\alpha, q_\alpha)\}$  and  $\mathcal{Y} = \{\sigma^{n_0}, 1\}$ . Applying the definition of the Methods section, it follows that the weights  $\mathbf{w}$  are in this case equal to the  $q_\alpha$ s. By averaging over all values of  $n_0$ , we now obtain

$$\langle \text{EMD}(\{(s^\alpha, q_\alpha)\}) \rangle_{n_0} = \sum_{n=1}^N \sum_{\alpha} H(s^\alpha, \sigma^n) \cdot q_\alpha.$$

By the same calculation procedure as in the previous section, this expression simplifies to

$$\langle \text{EMD}(\{(s^\alpha, q_\alpha)\}) \rangle_{n_0} = \sum_{i=1}^L \left( 1 - \sum_{a=1}^q p_i(a) q_i(a) \right),$$

where the profile of the present population  $p_i(a)$  has already been defined, and  $q_i(a)$  stands for the profile of the predictor, that is

$$q_i(a) = \sum_{\alpha} \delta_{s_i^\alpha, a} q_\alpha.$$

1035 To minimize this distance, we find a profile  $q_i(a)$  that maximizes the quantity  $\sum_{\alpha} \delta_{s_i^\alpha, a} q_\alpha$  for each position  $i$ . It is  
 1036 clear that this is done by assigning a value  $q_i(a) = 1$  if  $a$  maximizes  $p_i(a)$ , and  $q_i(a) = 0$  otherwise. Thus, the profile of  
 1037 the predictor must be that of the consensus sequence, which is only possible if the predictor becomes  $\{\sigma^{cons}, 1\}$ .

## 1038 3. Correcting for nested trajectories

1039 The analysis of the main text computes probabilities of fixation assuming that all trajectories are independent.  
 1040 However, it is well-known that mutations in influenza viruses are nested: they appear on backgrounds that already  
 1041 carry other mutations. Since mutations appearing on the same genomes will jointly fix or disappear, many frequency  
 1042 trajectories are not independent but correlated. In order to compensate for potential biases due to this effect, we  
 1043 attempted to cluster trajectories based on similarity in their strain composition. Our aim is that two trajectories  
 1044 corresponding to mutations appearing mostly on the same genomes will be grouped in the same cluster. We then  
 1045 conduct the same analysis as in the main text on a set of *effectively independent* trajectories constructed by taking  
 1046 one trajectory from each cluster.

In order to perform clustering, we define a distance between trajectories. A frequency trajectory  $X$  is characterized by a series of frequency values and a time interval  $T$ .  $f(t)$  for  $t \in T$  corresponds to the frequency at which a given mutation  $x$  appears in the population at date  $t$ , and we define  $S(t)$  as the strains that carry mutation  $x$  at date  $t$ . With this notation,  $f(t)$  is the ratio of the number of elements in  $S(t)$  to the total number of strains at date  $t$ . Let us now consider two frequency trajectories  $X_1$  and  $X_2$ .

We define the distance  $d(X_1, X_2)$  between these two trajectories based on the average similarity of the strains  $S_1$  and  $S_2$  that compose them:

$$d(X_1, X_2) = \frac{1}{|T_1 \cap T_2|} \sum_{t \in T_1 \cap T_2} \frac{|S_1(t) \cap S_2(t)|}{|S_1(t) \cup S_2(t)|},$$

where  $T_1 \cap T_2$  is the time interval where both trajectories are active, and  $|\cdot|$  denotes the number of elements of a set. The quantity summed corresponds to the Jaccard index between strains composing  $X_1$  and  $X_2$  at a given date. It is 1 if the two trajectories share exactly the same strains for this date, and 0 if they share no strain at all. This leads to the two following properties of  $d$ :

- if  $d(X_1, X_2) = 0$ , then  $X_1$  and  $X_2$  represent the same frequency trajectory. The mutations  $x_1$  and  $x_2$  that they correspond to always appear on the same strains and are totally linked.
- if  $d(X_1, X_2) = 1$ , then  $X_1$  and  $X_2$  can be considered completely independent. This can be the case if the two trajectories do not occur at the same dates, \*i.e.\*  $|T_1 \cap T_2| = 0$ , or if their respective mutations are never present on the same genomes.

we attempt to reduce the potential statistical bias due to the nesting of trajectories by grouping them based on the above defined distance. Given a set of trajectories  $\{X\}$ , we perform a decomposition of  $\{X\}$  into disjoint clusters  $C_1(d^*) \cup \dots \cup C_n(d^*) = \{X\}$  where  $d^*$  is an arbitrary threshold distance. Clusters are built in such a way that given two trajectories  $X_i$  and  $X_j$

$$d(X_i, X_j) \leq d^* \Rightarrow \exists k : X_i, X_j \in C_k$$

and

$$X_i \in C_k \Rightarrow \exists j \in C_k : d(X_i, X_j) \leq d^*.$$

These condition imply that the clusters formed are the minimal ones that guarantee that any two trajectories closer than the threshold distance  $d^*$  belong to the same cluster. The number of clusters  $n$  depends on the chosen value for  $d^*$ .

We compute clusters for different values of  $d^*$  for the case of the HA gene in A/H3N2. The top panel of figure S1 shows the number cluster  $n$  as a function of  $d^*$ . In the  $d^* = 0$  case, only trajectories that are exactly identical in terms of strain composition are clustered together. In this case, our clustering amounts to counting mutations that appear on exactly the same strains as one, reducing the number of effective trajectories from 800 to slightly less than 700. For higher values of  $d^*$ , the number of clusters steadily goes down until it reaches 1 for  $d^* = 1$ , which is the maximum value of the distance  $d(X_1, X_2)$ . The sharp drop in  $n$  for  $d^* = 0.5$  is explained by the high number of very short (typically one time point) and low frequency trajectories that share one out of two strains.

Since the choice of  $d^*$  is arbitrary and since no particular value can be chosen based on the number of clusters  $n(d^*)$ , we decide to test our clustering strategy for five values, namely  $d^* \in \{0, 0.05, 0.1, 0.2, 0.49\}$ . The bottom panel of figure S1 shows examples of a cluster for the four non-zero values of  $d^*$ . The cluster displayed in each case is the one containing the mutation HA1:33R. As  $d^*$  increases, more and more unlike trajectories are grouped together. In the case  $d^* = 0.49$ , the cluster consists of 13 trajectories, 5 of which end up dying while the rest fix. Since such a high value of  $d^*$  results in grouping trajectories that do not have the same fate (fixation or death), we decide to exclude it from the rest of the analysis, resulting in four remaining values  $d^* \in \{0, 0.05, 0.1, 0.2\}$ .

Once clustering is performed, we re-conduct the analysis of the main text on a set of effective trajectories. This set is constructed by taking one trajectory at random from each cluster. Effective trajectories are then considered independent from each other. The left panel of figure S2 shows the fixation probability of trajectories as a function of their frequency for different values of  $d^*$ , for the HA gene of A/H3N2. The result obtained in panel A of figure

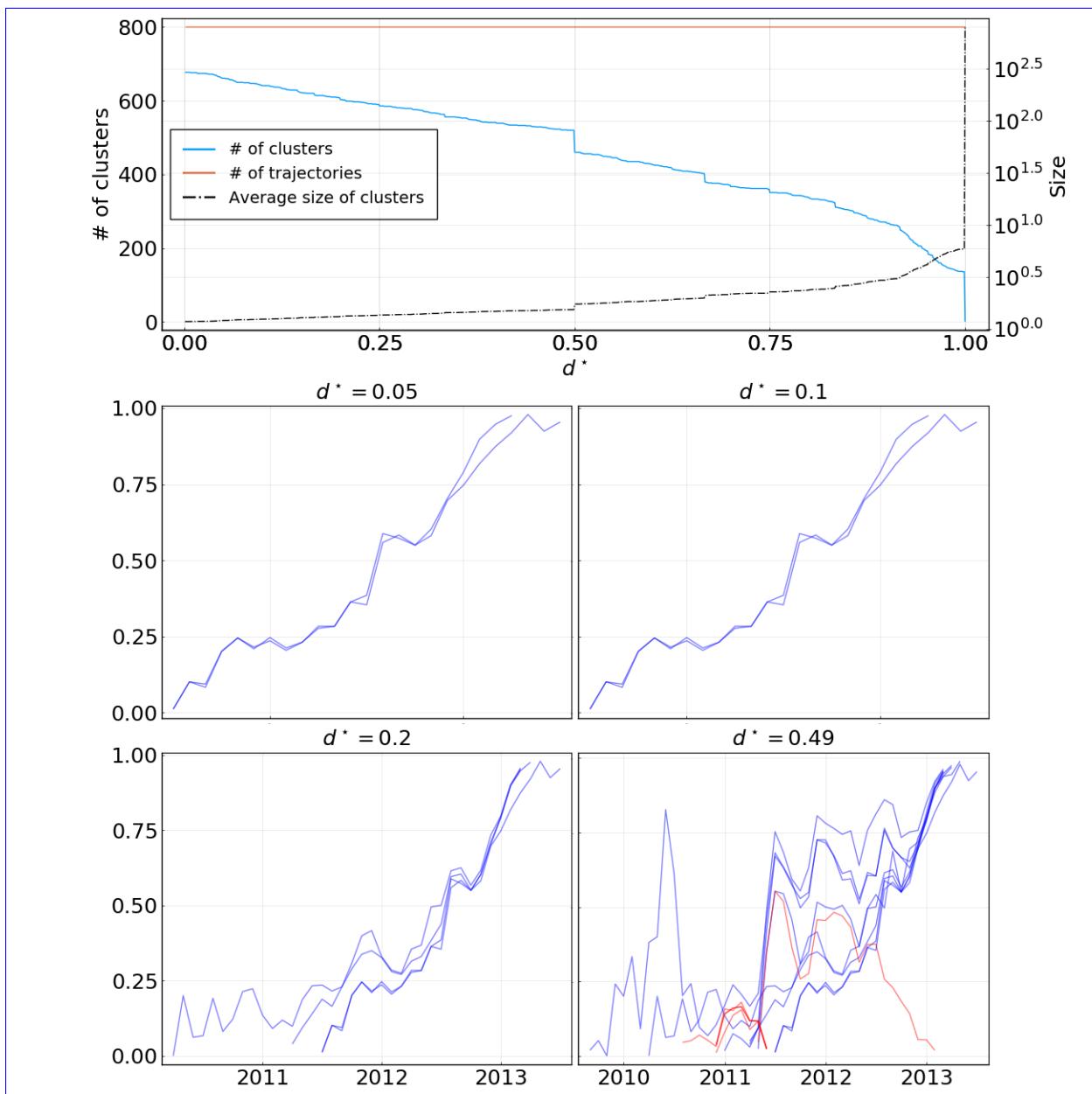


Figure S 1 **Top:** Left-axis: number  $n$  of clusters as a function of the threshold distance  $d^*$ . The total number of trajectories is shown as a flat orange line. Right-axis: average size of clusters as a function of  $d^*$ . **Bottom:** Examples of clusters for four values of  $d^*$ . The four clusters displayed are the ones to which the trajectory of mutation HA1:33R belongs.

2 of the main text is also showed as a reference. For the three lower values of  $d^*$ , results do not differ from the one obtained in the main text, even though the number of trajectories in each frequency bin has dropped as can be seen in the right panel of figure S2. This indicates that grouping together trajectories that share most of their strains, and are thus very correlated, does not modify the computed fixation probability in any way. For the higher value  $d^* = 0.2$ , fixation probability drops slightly across all frequency bins, suggesting that fixating trajectories tend to be grouped together more frequently. However, this drop remains of limited amplitude.

Overall, this analysis leads us to think that even though mutations in influenza may be nested, considering trajectories as independent does not result in strong statistical biases. Indeed, clustering similar trajectories together does significantly modify results presented in the main text.

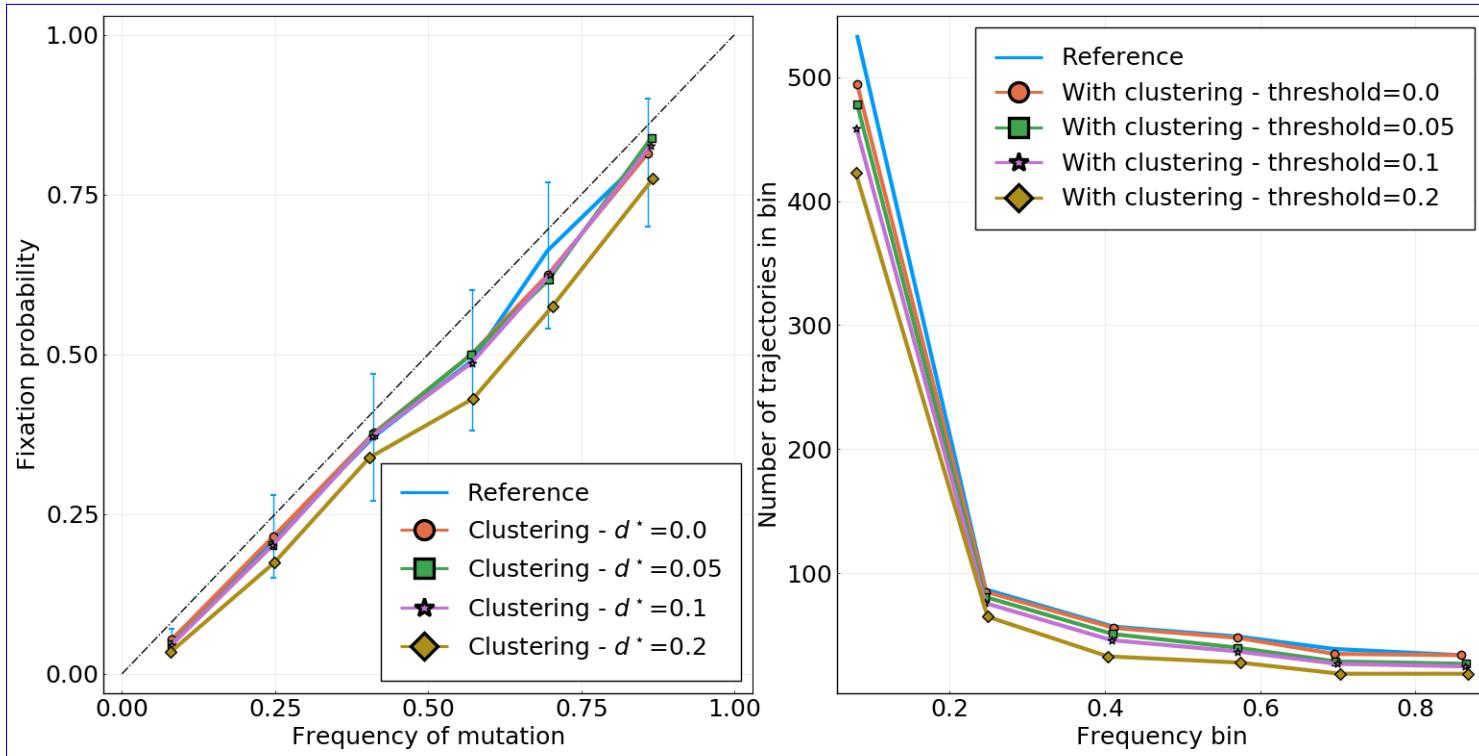


Figure S 2 **Left:** Fixation probability of trajectories as a function of probability for four values of  $d^*$ . The reference curve is the same as in panel A of figure 2 for the HA gene. For readability, error bars are only displayed for the reference curve. **Bottom:** Number of trajectories in each frequency bin corresponding to the left panel.

#### 1087 4. Ability of fitness models to predict fixation

1088 In a companion article (?), authors developed several models to predict the future HA population for A/H3N2  
 1089 influenza. The premise of these models is to assign a fitness score to the HA gene of each A/H3N2 strain in a  
 1090 given year, and to then apply deterministic evolutionary equations to obtain the future strain composition. The  
 1091 fitness values are computed using several quantities that have been shown to be connected with viral fitness. Fitness  
 1092 depends either on a single or on two quantities, and coefficient(s) of each fitness model are trained by minimizing the  
 1093 earth movers distance between the observed strain population one year in the future and the estimated population  
 1094 produced by the model. The training is performed using HA sequences from 1990 to 2015, and we consider here the  
 1095 models obtained by taking the average value of coefficients obtained for every year.

1096 We first assess the ability of the four most performing fitness models based on individual scores in (?) to  
 1097 predict fixation: mutational load (Łuksza and Lässig, 2014), hemagglutination inhibition (HI) antigenic novelty  
 1098 (Neher et al., 2016), a “delta frequency” score based on the recent increase in frequency of clades, and the previously  
 1099 mentioned Local Branching Index (LBI) (Neher et al., 2014). Figure S3 shows fixation probability for HA mutations  
 1100 with fitness scores in the top or bottom half of the fitness distribution for the four fitness measures.

1101 The best performing individual score (and second overall) in (?) LBI, does not provide any information about  
 1102 fixation (figure S3A). This is consistent with the result found in figure 3 of the main text. In panels B and D, we  
 1103 observe the same inability to predict fixation for the fitness models based on HI titer and delta frequency, which were  
 1104 respectively the second and third best individual scores (fifth and sixth overall) in (?). However, the score based  
 1105 on mutational load (panel C) does show a more significant predicting power, with higher fitness mutations having a  
 1106  $\sim 25\%$  higher chance of fixing than lower fitness ones for three frequency bins in a row. However, this effect vanishes  
 1107 for higher frequencies, with both fitter and less fit mutations having the same chance to fix.

1108 Next, we use two fitness models based on linear combinations of the following quantities: mutational load + LBI  
 1109 and mutational load + HI antigenic novelty. These composite models were the first and third best performing ones  
 1110 in (?). Figure S4 again shows fixation probability for HA mutations with fitness scores above or below the median  
 1111 fitness value.

1112 Both models provide some information about fixation, with the mutational load + HI model having more success

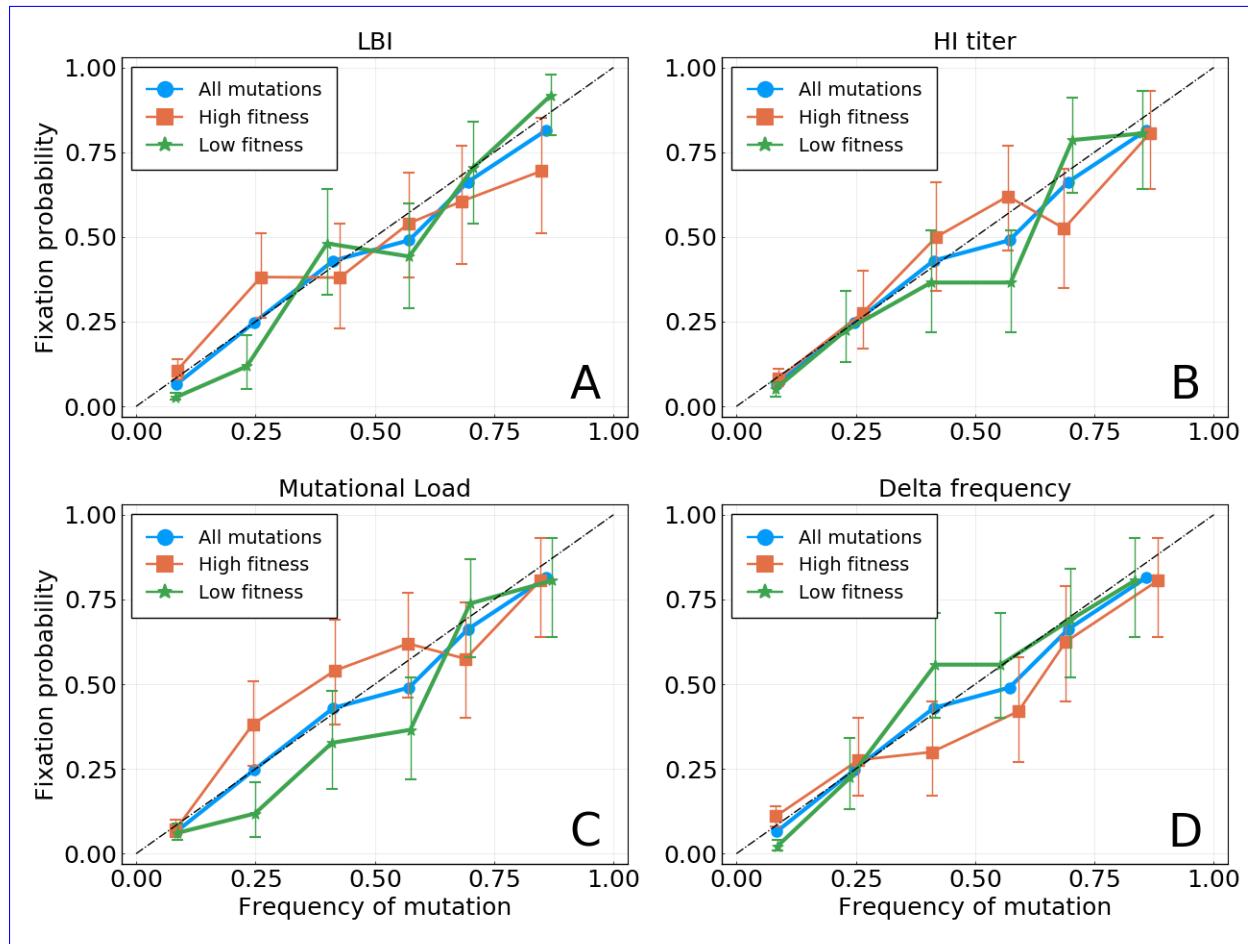


Figure S 3 Ability of fitness scores based on **A** LBI, **B** HI titer, **C** mutational load and **D** delta frequency, to predict fixation of mutation frequency trajectories in the HA gene of A/H3N2. Similar to figure 3 of the main text.

for frequencies around 0.5. This observation could be explained by the fact that fitness based on HI titers alone showed some power in this same frequency range. However, both models do not seem to perform significantly better than the one based on mutational load alone. This is not surprising, as LBI and HI titer did not perform well as individual models.

Combining models in (?) typically achieved a greater performance than using single models. This was the case for the mutational load + LBI model, combining the second and 7th best predictors overall to obtain the best one, and for the mutational load + HI titer model that combined the 4th and 7th best predictors to obtain the third best one. However, our results indicate that the gain in capability of predicting the future strain composition obtained by combining models is not immediately transferable to the exercise of predicting fixation.

It is important to note that the prediction targets in (?) and in the present article are not identical. It is shown in figure 4 of the main text that even though LBI performs well when predicting the future strain composition, it may not be because it is indicative of fitness.

## 5. Biases in frequency estimations

The frequency of mutations in a given time-bin is simply performed by computing their frequency in sequences sampled in that time bin. This leads to potential biases in estimating frequencies, that arise for two reasons:

- (i) A mutation present at frequency  $p$  in the population might be observed at another frequency  $f \neq p$  if  $f$  is estimated using a sub-sample of the population.
- (ii) For a neutrally evolving population, the distribution of frequencies of alleles is of the form  $P(p) \propto 1/p$ . This

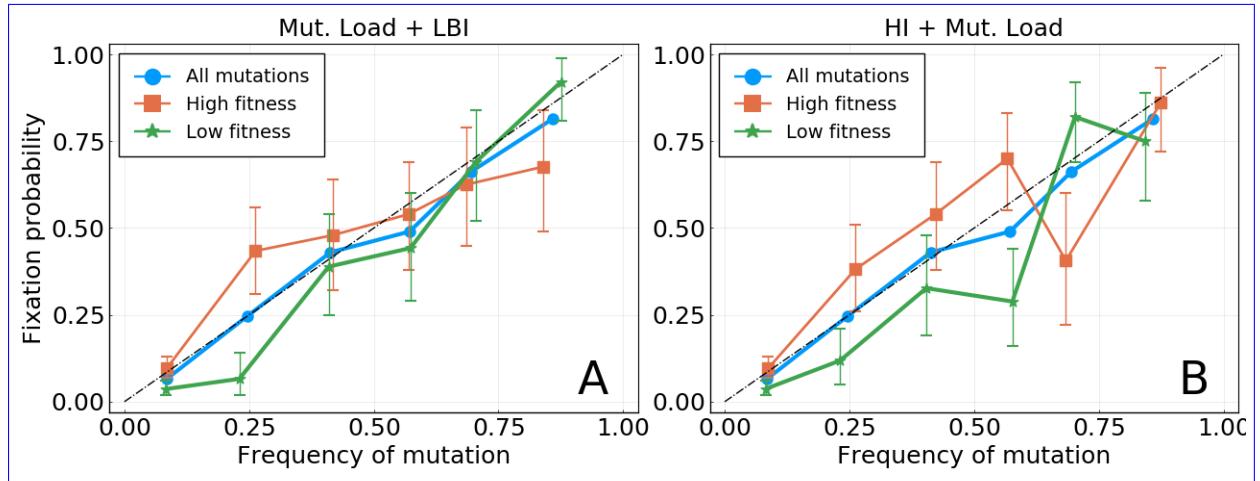


Figure S 4 Ability of fitness scores based on two combinations of the four fitness scores mentioned above: **A** mutational load + LBI; **B** mutational load + HI titer.

means that the amount of alleles at frequency  $p$  is lower when  $p$  is higher.

To illustrate (i), let us compute the probability that a mutation present at “real” frequency  $p$  in the population is found to be in a given frequency bin  $[f_1, f_2]$  when  $p$  is estimated from a sample of size  $n$ . The sample consists of  $n$  observations  $\{x_i\}$  with  $1 \leq i \leq n$ , with  $x_i = 1$  if sequence sequence  $i$  of the sample bears the mutation, and  $x_i = 0$  if not. If  $n$  is small with regard to the total population size, we can consider the  $x_i$  as random variables with a binomial distribution, meaning that  $P(x_i = 1) = p$  and  $P(x_i = 0) = 1 - p$ . The empirical frequency  $f$  is then estimated by taking the average of the  $x_i$  variables, that is  $f = (x_1 + \dots + x_n)/n$ . If those are independently sampled and  $n$  is large enough, the probability of measuring value  $f$  is given by the Central Limit Theorem:

$$P_{n,p}(f) \propto e^{(f-p)^2/2\sigma^2}, \text{ where } \sigma^2 = \frac{p(1-p)}{n}. \quad (5)$$

To compute the probability that this mutation is found in a given frequency bin  $[f_1, f_2]$ , we integrate this distribution:

$$P_{f_1, f_2}(p, n) = \int_{f_1}^{f_2} dx P_{n,p}(x). \quad (6)$$

Function  $P_{f_1, f_2}(p, n)$  is shown as a function of  $p$  for a fixed interval and for different values of  $n$  in the first panel of figure S5. Note the asymmetry of it: the variance of a binomial distribution of parameter  $p$  is small when  $p$  is close to 0 or 1, and goes through a maximum at  $p = 0.5$ . For this reason, mutations present at frequency  $p$  close to 0.5 have a higher probability of being observed in other frequency bins. On the contrary, this is unlikely for very rare or very frequent mutations.

We now try to estimate biases in frequency estimation due this phenomenon. Given a set of mutations that have been measured in frequency bin  $[f_1, f_2]$ , what is the average *real* frequency of these mutations? To compute this, we need to sum  $P_{f_1, f_2}(p, n)$  over all possible real frequencies  $p$ , giving us the amount of mutations that are observed in interval  $[f_1, f_2]$ , and weigh this sum by the frequency value  $p$  as well as by the background distribution of frequencies  $P_b(p) \propto 1/p$ . This last quantity represents the expected amount of mutations that are present at frequency  $p$  in the population. Note that there is no divergence problem as the smallest non zero frequency is  $1/N$ , where  $N$  is the population size. This leads us to the following expression for the average of “real” frequencies:

$$\begin{aligned} \langle p \rangle(f_1, f_2, n) &= \int_{1/N}^{1-1/N} dp P_{f_1, f_2}(p, n) P_b(p) p \\ &= \int_{1/N}^{1-1/N} dp P_{f_1, f_2}(p, n). \end{aligned} \quad (7)$$

We have not made normalization explicit in these equations. It is simply achieved by dividing the above expression by  $\int dp P_{f_1, f_2}(p, n) P_b(p)$ .

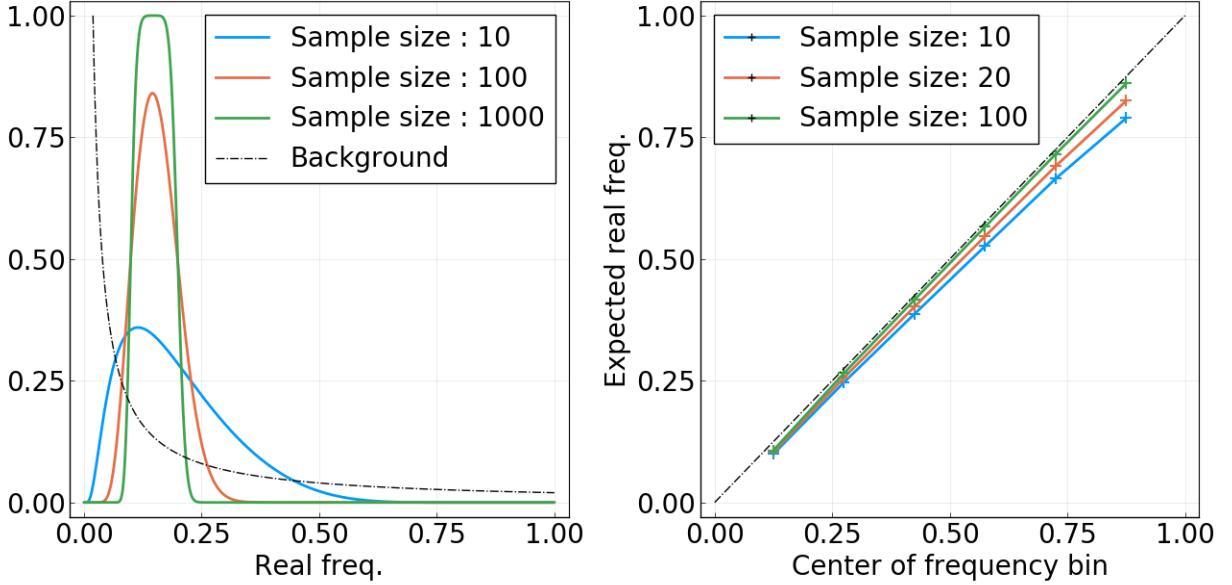


Figure S 5 **Left:** For a mutation present at frequency  $p$  in the population, probability of being observed in the frequency bin  $[0.1, 0.2]$  as a function of  $p$  and for different sample sizes  $n$ . The dashed black line sketches the (non-normalized) background distribution  $P_b(p)$ . **Right:** Expected “real” average frequency of mutations found in frequency bin  $[f_1, f_2]$  as a function of the centre of the bin  $(f_1 + f_2)/2$ , for different sample sizes.

In the second panel of figure S5,  $\langle p \rangle(f_1, f_2, n)$  is plotted as a function of the centre of the interval  $[f_1, f_2]$  and for different values of  $n$ . For sample sizes  $n > 100$ , the biases due to this effect are almost non existent. For smaller samples, for instance  $n = 10$ , they are small but non negligible. However, we argue that this is not a significant problem with respect to the main results presented in this article. First, figure S10 shows that sample sizes of the order of  $n = 10$  are only the case for a few months in the period going from year 2000 to 2018. From 2010 and onwards, more than a hundred sequences are available per month for most months. Secondly, even if most samples were in the  $n = 10$  case, deviations shown in figure S5 are small enough that results shown in figures 2 and 3 would be *qualitatively* unchanged.

Note that using the centre of the interval as a reference in figure S5, *i.e.*  $(f_1 + f_2)/2$ , would be correct in the case of a very large  $n$  and a flat background distribution  $P_b(p)$ . For figures 2 and 3 of the main text however, the average frequency of mutations found in an interval  $[f_1, f_2]$  is computed by taking the average of the observed frequencies, and not the centre of the interval. This partially takes into account biases considered here, as the background distribution  $P_b(p)$  is then accounted for, even though it is equivalent to assuming infinite sample sizes.

## 1169 6. Cutting off the HA1 159S branch

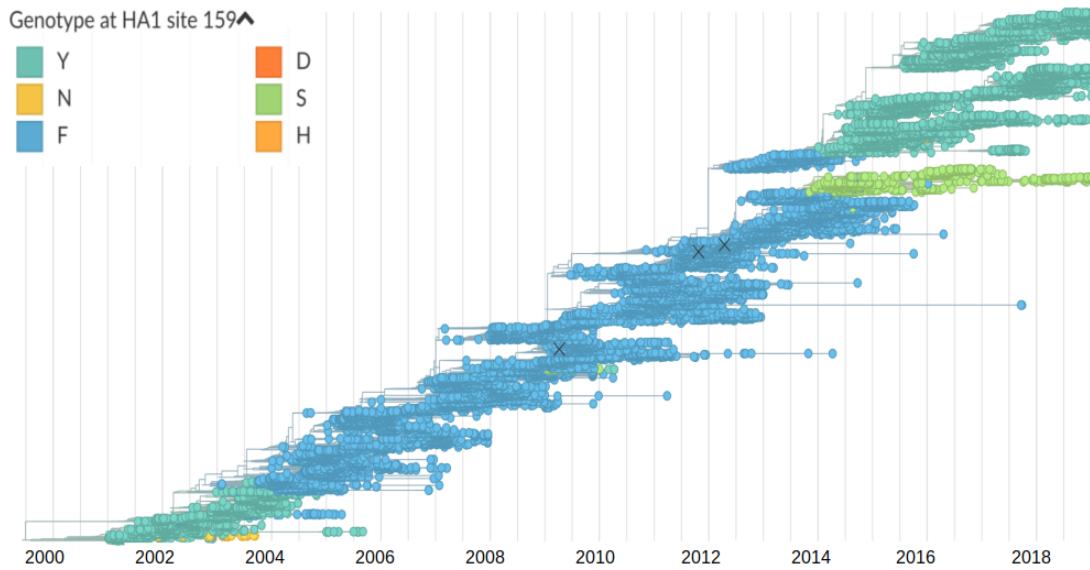


Figure S 6 Tree used for this study, based on a random selection of 100 strains per month from year 2002 to 2018. Nodes and branches are colored according to the amino acid found at position HA1:159. The HA1 159S mutation is visible as a thin but long light-greened color branch, coalescing with the “trunk” around 2013.

1170 The analysis of the main text is in a large part based on the probability of fixation of mutations. The motivation  
 1171 underlying this choice is the relatively short coalescence time of the A/H3N2 influenza population, typically around  
 1172 three years. This can be seen in figure 2 of the main text, which shows the typical lifetime of frequency trajectories,  
 1173 ending in fixation or loss after at most 3 years in most cases. The tree in figure S6 is another illustration of this: for  
 1174 the most part of it, a “trunk” is clearly identifiable, and lineages that depart from it have a relatively short lifetime.  
 1175 This is no longer the case since the year ~ 2013: two clades have been competing since then, with no definite way to  
 1176 identify a trunk in the tree. The clade defined by the HA1 159S mutation, colored in light green on figure S6, is one of  
 1177 these two competing lineages. Because of this particular situation, the number of mutations fixating in the population  
 1178 is strongly reduced, as a mutation must appear in both clades to reach a frequency of 1. This is a potential flaw in our  
 1179 analysis, which concentrates on mutations fixating.

1180 For this reason, we decided to re-run our analysis after having cut off the HA1 159S clade. In other words, we remove  
 1181 from the set of sequences those that carry the HA1 159S mutation. Results are shown in figures , equivalent to figures 2  
 1182 and 3 of the main text. It is clear that qualitative results are left unchanged when this competing clade is removed.  
 1183 This can be surprising, as almost no complete fixation of an amino acid mutation has occurred since 2013. Cutting off  
 1184 the HA1 159S branch should thus result in many new fixations, changing the analysis. The reason for the similarity  
 1185 of results can be explained: fixation (resp. loss) of a mutation are defined here as the frequency of this mutation  
 1186 being measured above 95% (resp. 5%) frequency for two months in a row. As the HA1 159S clade is rather sparsely  
 1187 populated, it reaches frequencies lower than 5% two times (in 2015 and 2017), allowing mutations in the competing  
 1188 clade to “fix” as defined here. Thus, removing strains carrying HA1 159S does not introduce a significant amount of  
 1189 “new” fixation events.

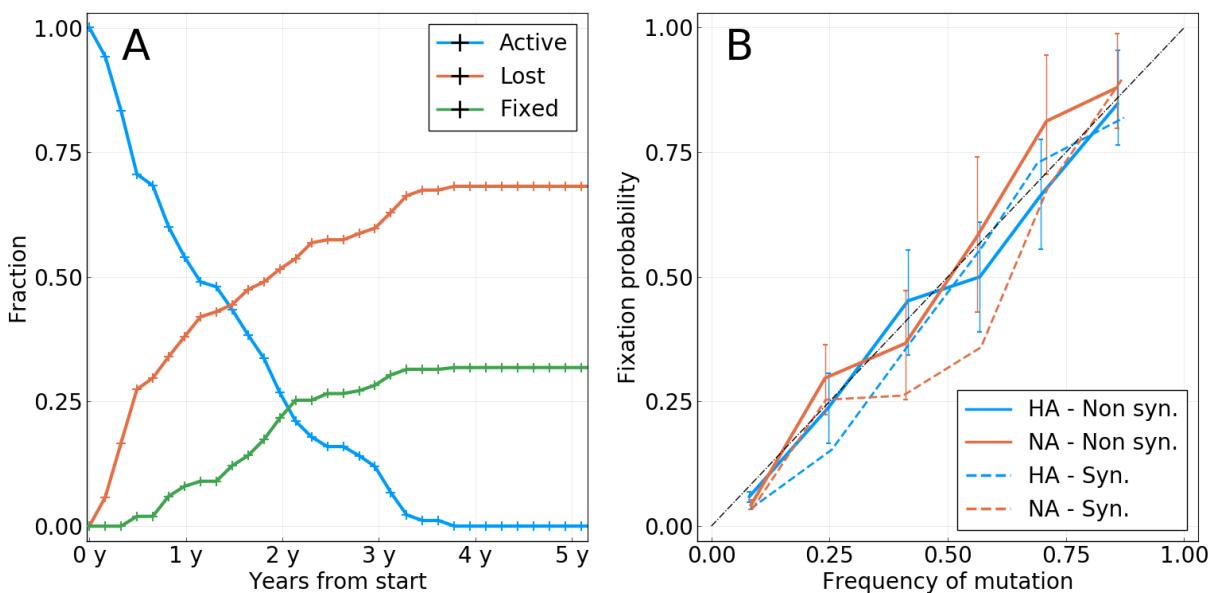


Figure S 7 Equivalent to figure 2 of the main text, but with strains carrying the HA1 159S mutation removed.

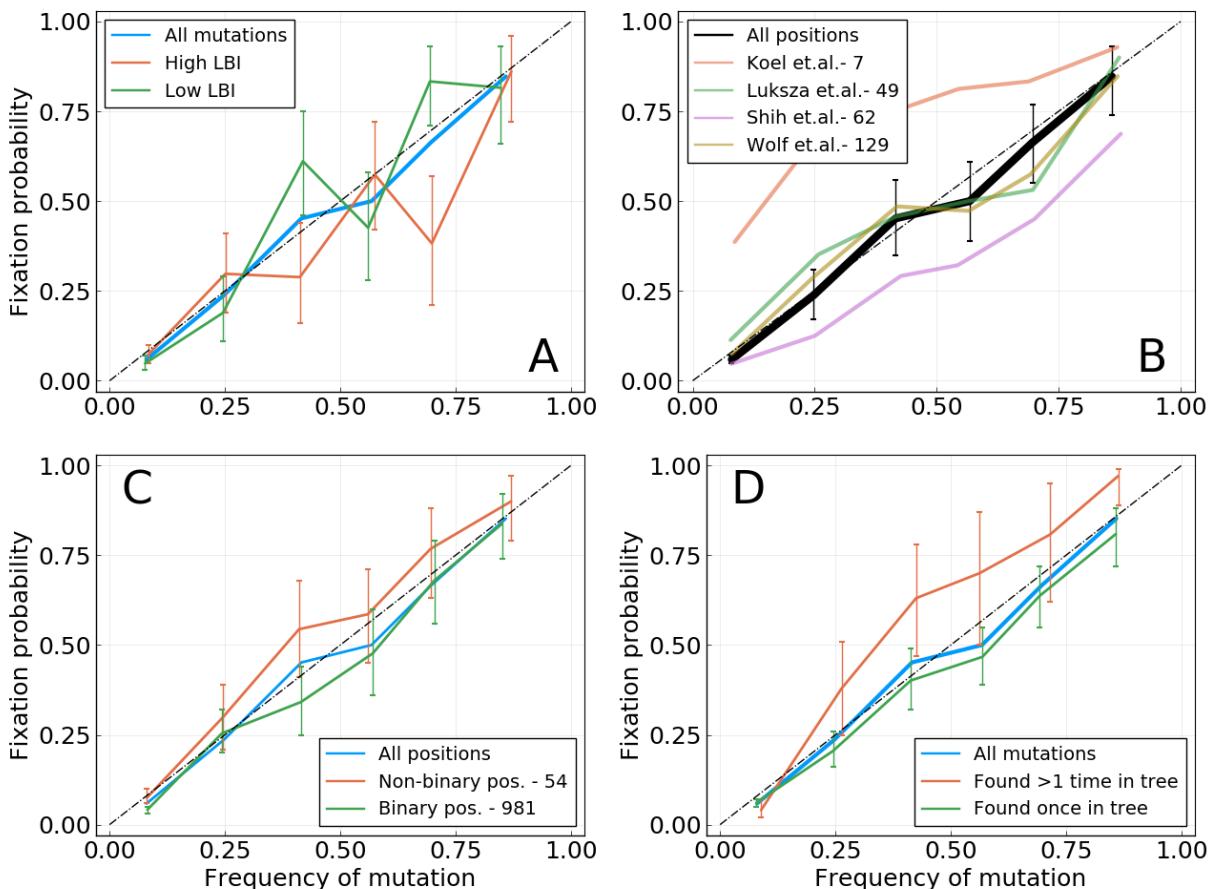


Figure S 8 Equivalent to figure 3 of the main text, but with strains carrying the HA1 159S mutation removed.

1190 **7. Probability of fixation in single locus model of evolution**

1191 In (Kimura, 1964), Kimura investigates a simple model of evolution with a single locus and a population of size  
 1192  $N$ . In this framework, a mutation at this locus with fitness effect  $s$  and observed at frequency  $f$  has the following  
 1193 probability of fixation:

$$P_{fix}(f|s, N) = \frac{1 - e^{-sNf}}{1 - e^{-sN}}. \quad (8)$$

1194 Expanding this formula for  $sN \ll 1$ , that is in the weak selection regime, yields at the first order

$$P_{fix}(f|s, N) = f + f(1 - f) \frac{sN}{2}. \quad (9)$$

1195 Equation 9 tells us two things. First, when the mutation is neutral, that is  $s = 0$ , we have  $P_{fix}(f) = f$ . This naturally  
 1196 confirms the result obtained for a neutral model of evolution. Seconds, when  $sN \neq 0$ , we can expect deviations from  
 1197 the diagonal in a  $P_{fix}$  against  $f$  plot. The sign of these deviations is determined by the sign of  $s$ , with beneficial  
 1198 mutations being found above diagonal while deleterious one are found below. The amplitude of these deviations  
 1199 depends on the strength of selection  $sN$ , as well as on the frequency through the  $f(1 - f)$  term, making them larger  
 1200 for  $f \sim 0.5$ .

## 1201 8. Mutation tables

Gene	Position	AA	Start date	End date	Shih	Luksza	Koel	Tree counts
HA1	144	D	2001-06-09	2002-02-04	true	true	false	0
HA1	189	N	2003-07-29	2004-05-24	false	true	true	2
HA1	159	F	2003-08-28	2004-05-24	false	true	true	2
HA1	226	I	2003-09-27	2004-09-21	true	true	false	3
HA1	145	N	2003-12-26	2004-11-20	false	true	true	2
HA1	227	P	2003-05-30	2005-04-19	false	true	false	2
HA2	32	I	2004-06-23	2005-07-18	false	false	false	1
HA1	193	F	2004-12-20	2006-03-15	false	true	true	1
HA2	46	D	2006-06-13	2007-05-09	false	false	false	2
HA2	121	K	2006-06-13	2007-06-08	false	false	false	1
HA1	50	E	2006-09-11	2007-06-08	false	true	false	2
HA1	140	I	2006-11-10	2007-11-05	true	false	false	1
HA1	173	Q	2007-07-08	2009-01-28	true	true	false	2
HA2	32	R	2007-07-08	2009-01-28	false	false	false	1
HA1	158	N	2009-01-28	2009-07-27	true	true	true	2
HA1	189	K	2009-01-28	2009-07-27	false	true	true	2
HA1	212	A	2009-03-29	2011-01-18	false	false	false	2
HA1	45	N	2010-03-24	2013-02-06	false	false	false	3
HA1	223	I	2010-12-19	2013-02-06	false	false	false	2
HA1	48	I	2011-03-19	2013-02-06	false	false	false	1
HA1	198	S	2011-03-19	2013-02-06	false	false	false	1
HA1	312	S	2009-08-26	2013-03-08	false	false	false	3
HA1	278	K	2011-06-17	2013-03-08	false	true	false	1
HA1	145	S	2011-04-18	2013-04-07	false	true	true	4
HA1	33	R	2011-06-17	2013-06-06	false	false	false	2
HA2	160	N	2012-07-11	2015-09-24	false	false	false	3
HA1	225	D	2013-08-05	2015-09-24	false	false	false	3
HA1	3	I	2013-08-05	2016-11-17	false	false	false	2
HA1	159	Y	2014-02-01	2016-11-17	false	true	true	2
HA1	160	T	2014-01-02	2017-07-15	false	true	false	2

Table S I The 30 trajectories that took place between year 2000 and year 2018 and resulted in fixation. Columns **Shih**, **Luksza** and **Koel** respectively indicate whether the position is found in the epitopes lists in (respectively) (Shih et al., 2007), (Luksza and Lässig, 2014) and (Koel et al., 2013). The **Tree counts** column indicates the number of times the mutation corresponding to the trajectory can be found in the phylogenetic tree. Note that a trajectory is only shown in the table if the sequenced population counts more than 10 strains at its time of fixation. This explains that only 30 trajectories are displayed, whereas more mutations did fix in this period of time.

Gene	Position	AA	Start date	End date	Fixation	Max. freq.
HA1	106	A	2001-02-09	2002-02-04	lost	1.0
HA1	144	D	2001-06-09	2002-02-04	fixed	1.0
HA1	105	H	2003-04-30	2003-10-27	lost	1.0
HA1	126	D	2003-04-30	2004-05-24	lost	1.0
HA1	140	Q	2004-01-25	2004-06-23	lost	0.31
HA1	226	I	2003-09-27	2004-09-21	fixed	1.0
HA1	173	E	2004-12-20	2006-03-15	lost	0.63
HA1	142	G	2006-06-13	2007-05-09	lost	0.71
HA1	144	D	2006-07-13	2007-05-09	lost	0.67
HA1	128	A	2006-09-11	2007-05-09	lost	0.25
HA1	157	S	2006-09-11	2007-05-09	lost	0.59
HA1	140	I	2006-11-10	2007-11-05	fixed	1.0
HA1	173	N	2007-12-05	2008-07-02	lost	0.3
HA1	157	S	2007-12-05	2008-09-30	lost	0.31
HA1	173	E	2006-06-13	2008-12-29	lost	0.67
HA1	173	Q	2007-07-08	2009-01-28	fixed	0.96
HA1	158	N	2009-01-28	2009-07-27	fixed	0.96
HA1	62	K	2009-01-28	2011-05-18	lost	0.73
HA1	144	K	2009-01-28	2011-05-18	lost	0.75
HA1	62	V	2011-04-18	2011-09-15	lost	0.34
HA1	157	S	2013-05-07	2015-09-24	lost	0.35
HA1	128	A	2012-08-10	2016-11-17	lost	0.81
HA1	197	K	2015-11-23	2016-11-17	lost	0.27
HA1	142	R	2018-05-11	2018-10-08	lost	0.38
HA1	142	G	2012-03-13		poly	0.86
HA1	144	S	2013-12-03		poly	0.96
HA1	121	K	2015-12-23		poly	0.82
HA1	142	K	2016-05-21		poly	0.77
HA1	62	G	2017-03-17		poly	0.75
HA1	128	A	2018-01-11		poly	0.56

Table S II Trajectories of mutations at epitope positions in (Shih et al., 2007) (*Shih et. al.*) that have been observed at least once above frequency 0.25. The **Fixation** column indicates whether the mutation has fixed, disappeared, or is still polymorphic as of October 2018. The **Max.freq.** column indicates the maximum frequency reached by the trajectory. A maximum frequency of 1 for mutations that finally disappear is explained by trajectories reaching frequency 1 for one time bin and going back to lower values for following ones (a frequency above 0.95 for two time bins in a row defines fixation).

Gene	Position	AA	Start date	End date	Fixation	Max. freq.
HA1	50	G	2001-02-09	2002-02-04	lost	1.0
HA1	144	D	2001-06-09	2002-02-04	fixed	1.0
HA1	126	D	2003-04-30	2004-05-24	lost	1.0
HA1	189	N	2003-07-29	2004-05-24	fixed	1.0
HA1	159	F	2003-08-28	2004-05-24	fixed	1.0
HA1	226	I	2003-09-27	2004-09-21	fixed	1.0
HA1	145	N	2003-12-26	2004-11-20	fixed	1.0
HA1	188	N	2004-07-23	2005-02-18	lost	0.36
HA1	227	P	2003-05-30	2005-04-19	fixed	1.0
HA1	173	E	2004-12-20	2006-03-15	lost	0.63
HA1	193	F	2004-12-20	2006-03-15	fixed	0.97
HA1	142	G	2006-06-13	2007-05-09	lost	0.71
HA1	144	D	2006-07-13	2007-05-09	lost	0.67
HA1	157	S	2006-09-11	2007-05-09	lost	0.59
HA1	50	E	2006-09-11	2007-06-08	fixed	0.95
HA1	173	N	2007-12-05	2008-07-02	lost	0.3
HA1	157	S	2007-12-05	2008-09-30	lost	0.31
HA1	173	E	2006-06-13	2008-12-29	lost	0.67
HA1	173	Q	2007-07-08	2009-01-28	fixed	0.96
HA1	158	N	2009-01-28	2009-07-27	fixed	0.96
HA1	189	K	2009-01-28	2009-07-27	fixed	0.96
HA1	213	A	2009-01-28	2010-02-22	lost	0.68
HA1	144	K	2009-01-28	2011-05-18	lost	0.75
HA1	53	N	2009-11-24	2013-02-06	lost	0.72
HA1	278	K	2011-06-17	2013-03-08	fixed	0.98
HA1	145	S	2011-04-18	2013-04-07	fixed	0.99
HA1	159	S	2013-11-03	2015-08-25	lost	0.46
HA1	157	S	2013-05-07	2015-09-24	lost	0.35
HA1	159	Y	2014-02-01	2016-11-17	fixed	0.97
HA1	159	S	2015-10-24	2016-11-17	lost	0.4
HA1	197	K	2015-11-23	2016-11-17	lost	0.27
HA1	160	T	2014-01-02	2017-07-15	fixed	0.96
HA1	142	R	2018-05-11	2018-10-08	lost	0.38
HA1	135	N	2018-06-10	2018-10-08	lost	0.38
HA1	142	G	2012-03-13		poly	0.86
HA1	144	S	2013-12-03		poly	0.96
HA1	121	K	2015-12-23		poly	0.82
HA1	142	K	2016-05-21		poly	0.77
HA1	131	K	2016-09-18		poly	0.77
HA1	135	K	2016-11-17		poly	0.47

Table S III Same as table SII, for (Luksza and Lässig, 2014) (*Luksza et. al.*).

Gene	Position	AA	Start date	End date	Fixation	Max. freq.
HA1	189	N	2003-07-29	2004-05-24	fixed	1.0
HA1	159	F	2003-08-28	2004-05-24	fixed	1.0
HA1	145	N	2003-12-26	2004-11-20	fixed	1.0
HA1	193	F	2004-12-20	2006-03-15	fixed	0.97
HA1	158	N	2009-01-28	2009-07-27	fixed	0.96
HA1	189	K	2009-01-28	2009-07-27	fixed	0.96
HA1	145	S	2011-04-18	2013-04-07	fixed	0.99
HA1	159	S	2013-11-03	2015-08-25	lost	0.46
HA1	159	Y	2014-02-01	2016-11-17	fixed	0.97
HA1	159	S	2015-10-24	2016-11-17	lost	0.4

Table S IV Same as table SII, for (Koel et al., 2013) (*Koel et. al.*).

## 1202 9. Supplementary figures

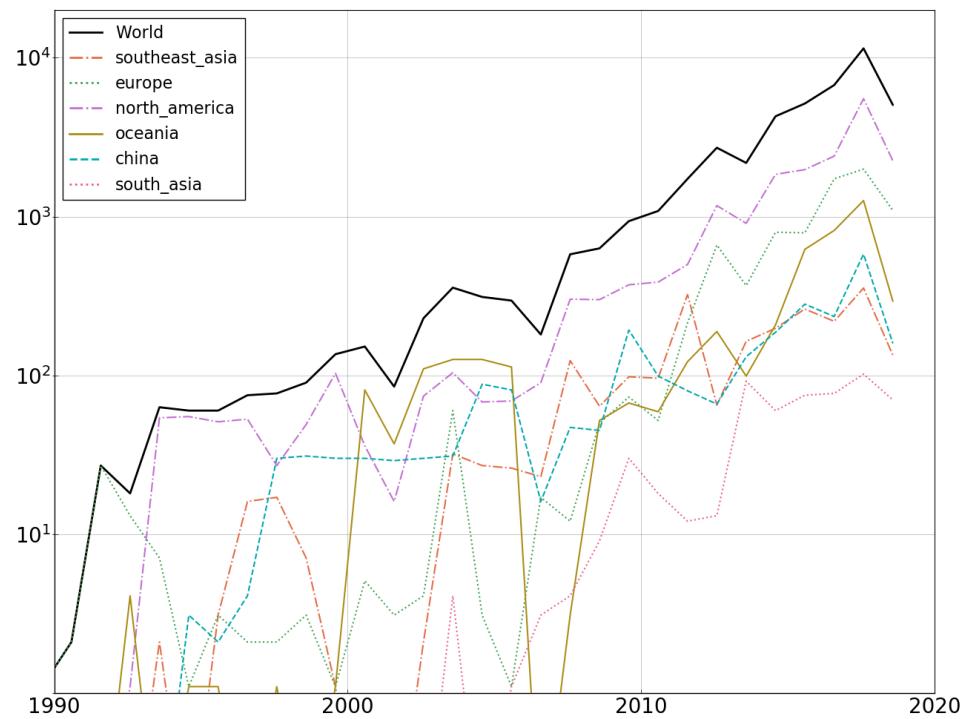


Figure S 9 Number of A/H3N2 HA sequences per year from year 1990.

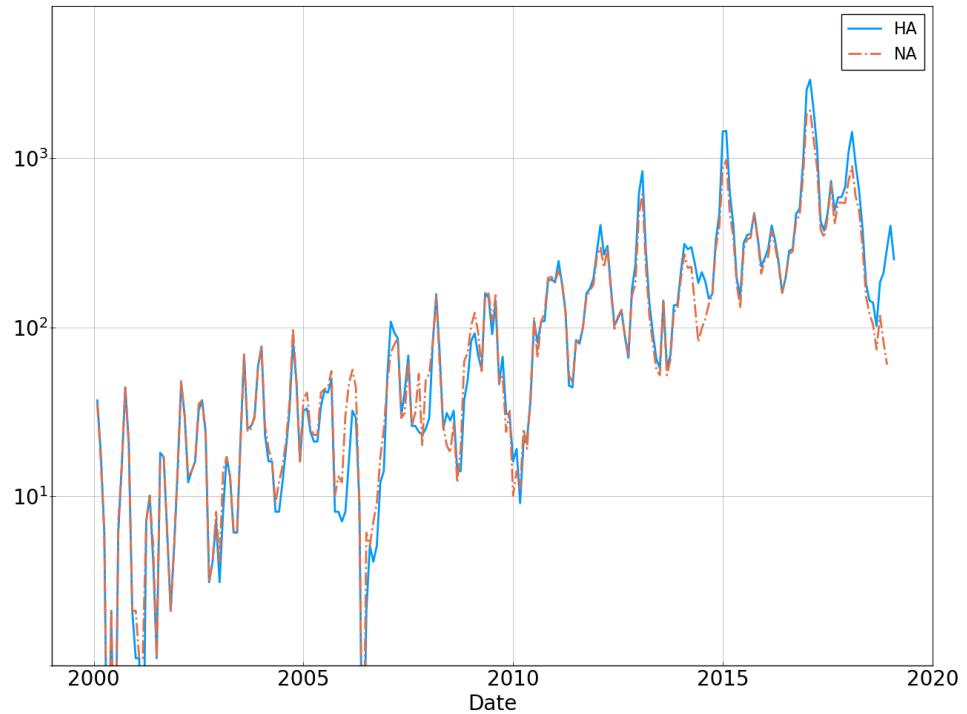


Figure S 10 Number of H3N2 HA and NA sequences per month from year 2000.

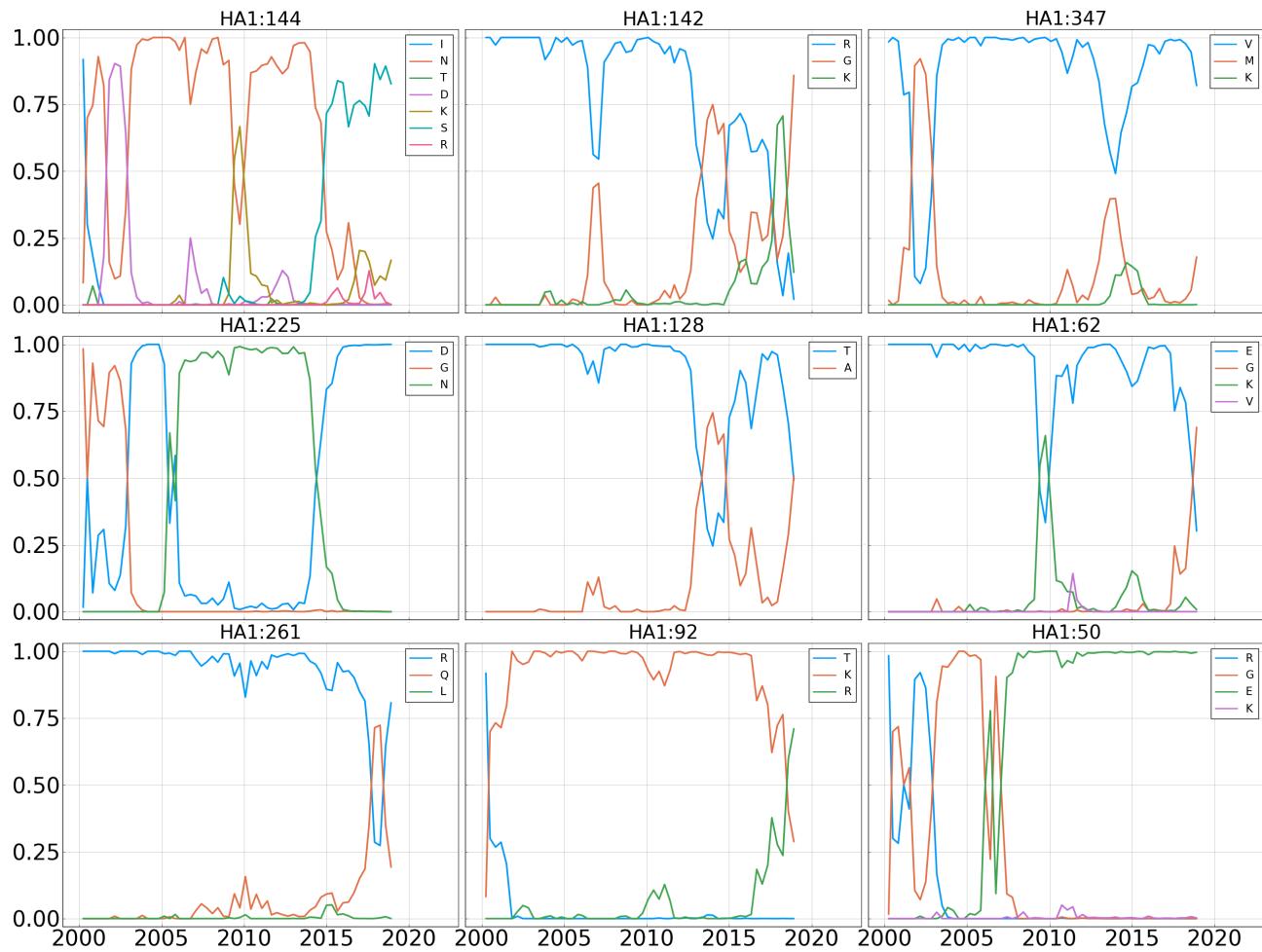


Figure S 11 Frequency trajectories for the 9 most entropic positions in the A/H3N2 HA protein.

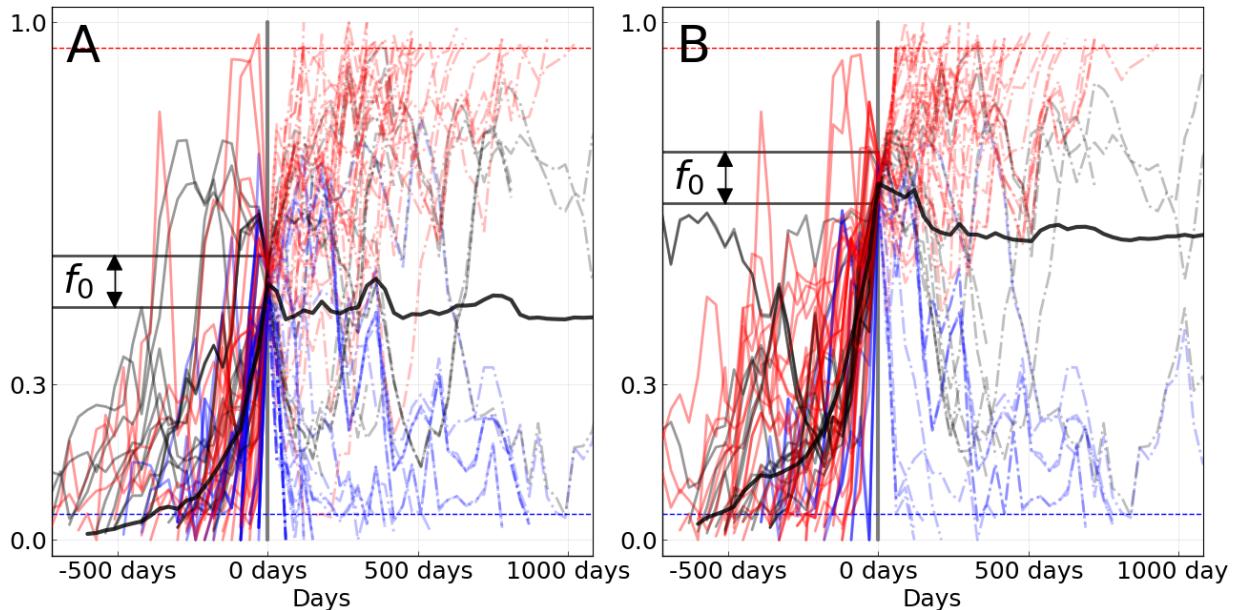


Figure S 12 Equivalent to panel B of figure 1 of the main text for A/H3N2, with  $f_0$  equal to 0.5 in **A** (76 trajectories), and 0.7 in **B** (63 trajectories).

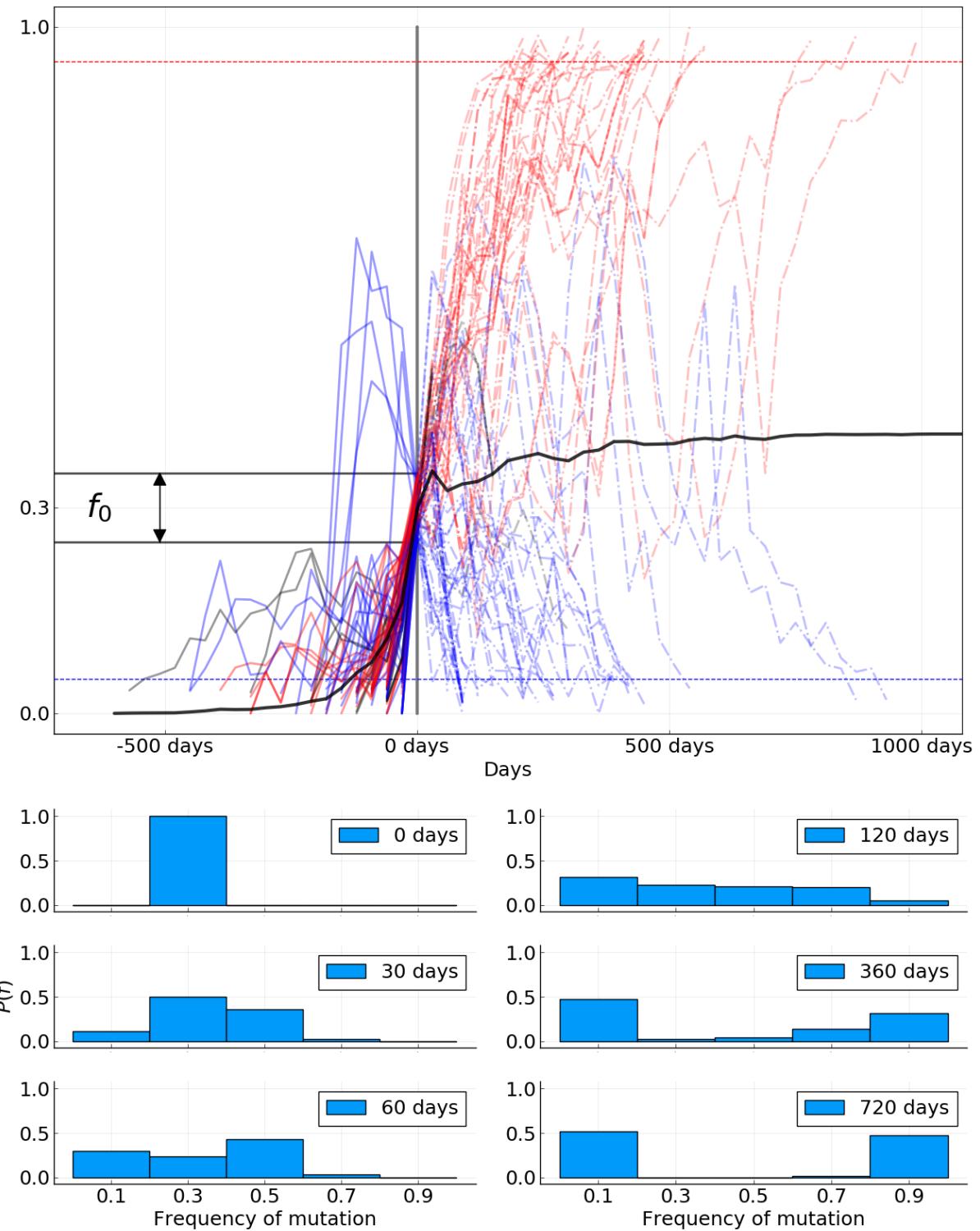


Figure S 13 Equivalent to panels **B** and **C** of figure 1 of the main text for A/H1N1pdm influenza. 89 trajectories are shown and participate to the mean (thick black line).

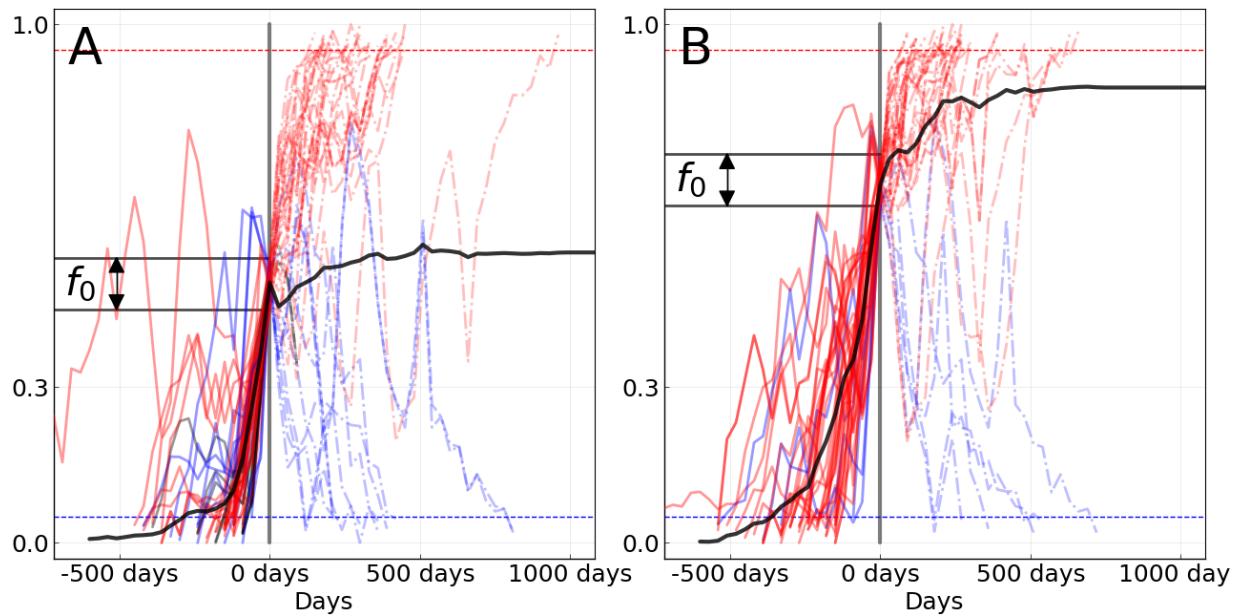


Figure S 14 Equivalent to panel **B** of figure 1 of the main text for A/H1N1pdm, with  $f_0$  equal 0.5 in **A** (50 trajectories), and 0.7 in **B** (41 trajectories).

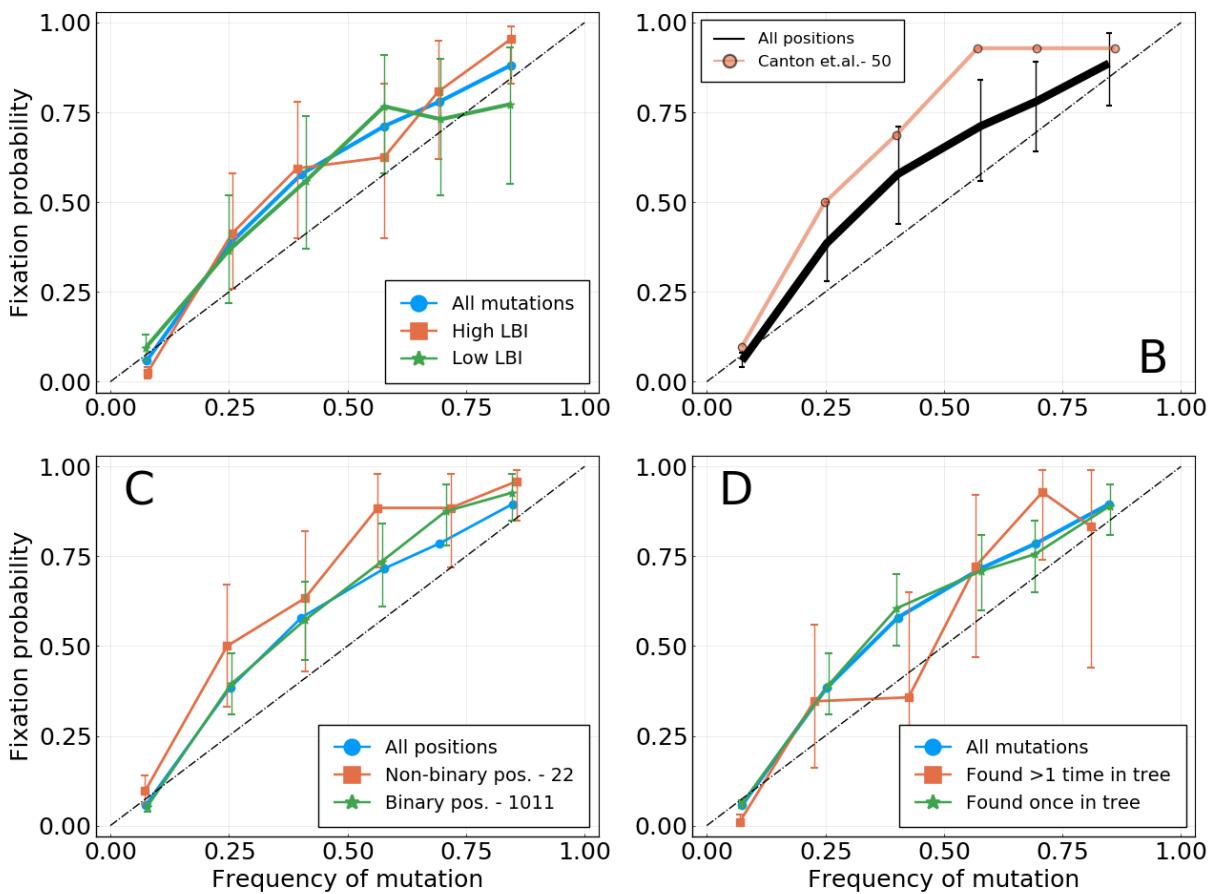


Figure S 15 Equivalent of figure 3 of the main text for the HA gene of A/H1N1pdm influenza. Fixation probability  $P_{fix}(f)$  as a function of frequency. **A:** Mutation with higher or lower LBI values, based on their position with respect to the median LBI value. **B:** Different lists of epitope positions in the HA protein. The authors and the number of positions is indicated in the legend. **C:** Mutations for binary positions, *i.e.* positions for which we never see more than two amino acids in the same time bin. **D:** Mutations that appear once or more than once in the tree for a given time bin.

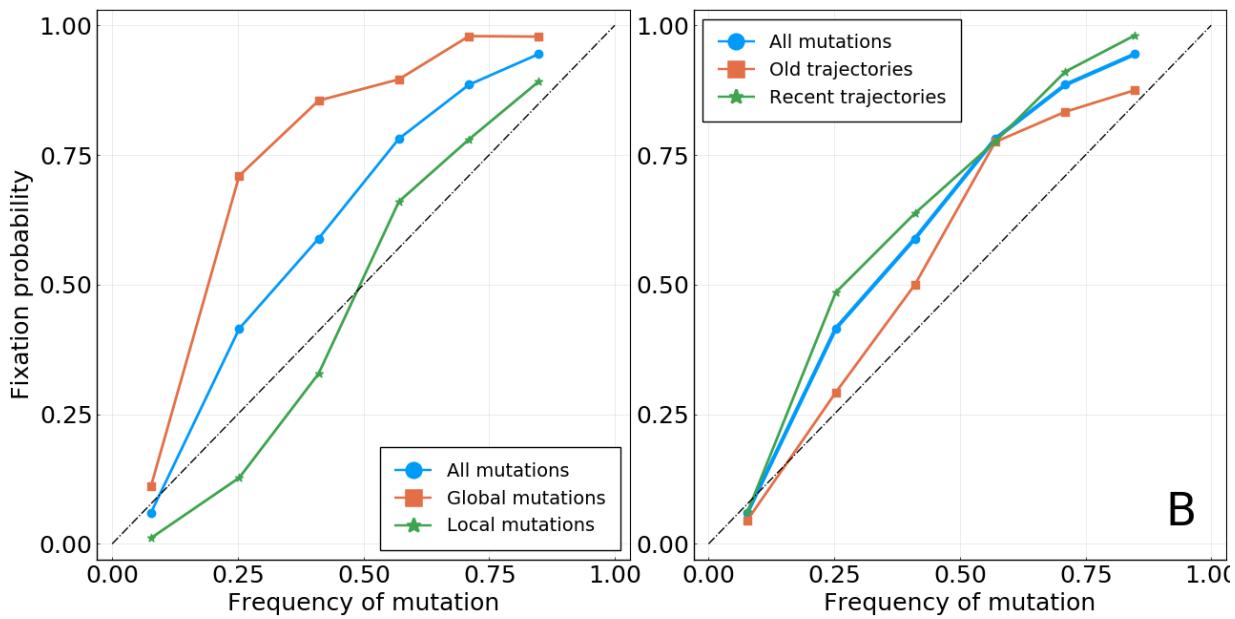
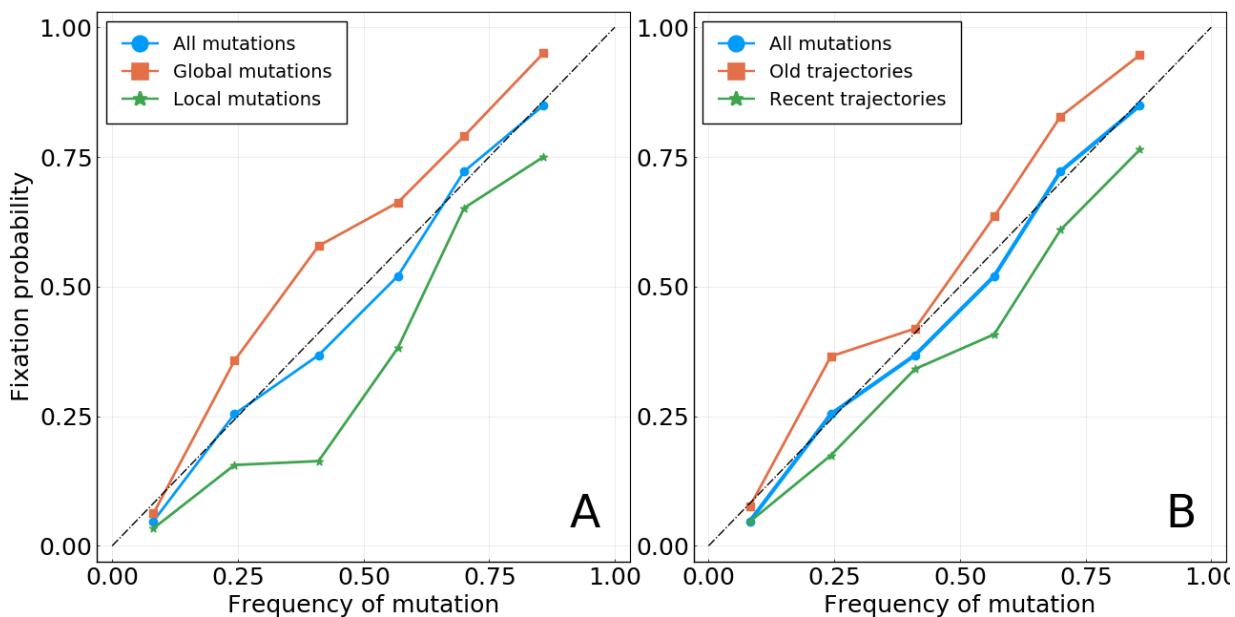


Figure S 17 Based on A/H1N1pdm HA and NA. **A:** Mutations with a higher or lower geographical spread, based on the median value of the score used (see Methods). Note: the words *local* and *global* only reflect the position of the geographic spread of the mutation relative to the median value computed for all mutations found at this frequency. As this median value may change with the considered frequency bin, so does the definition of local and global mutations. **B:** Mutations whose trajectories are older or more recent, based on the median age of trajectories when reaching the considered frequency  $f$ .

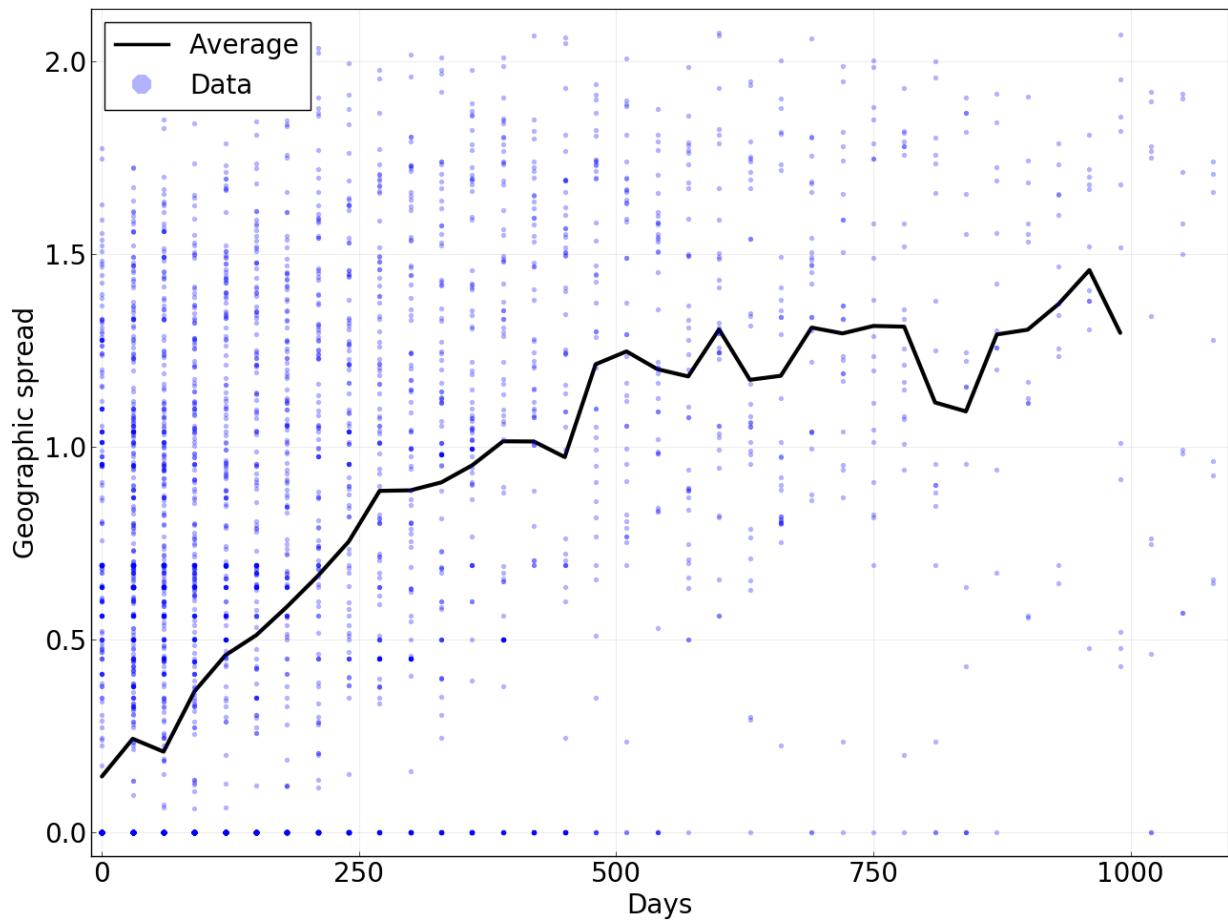


Figure S 18 Geographic spread of mutations as a function of the time for which they have been present in the population above a frequency of 5%. Points represent individual mutations and for a population in a given time bin. The line is the average of dots for a given value on the  $x$ -axis. Based on data for A/H3N2 HA.

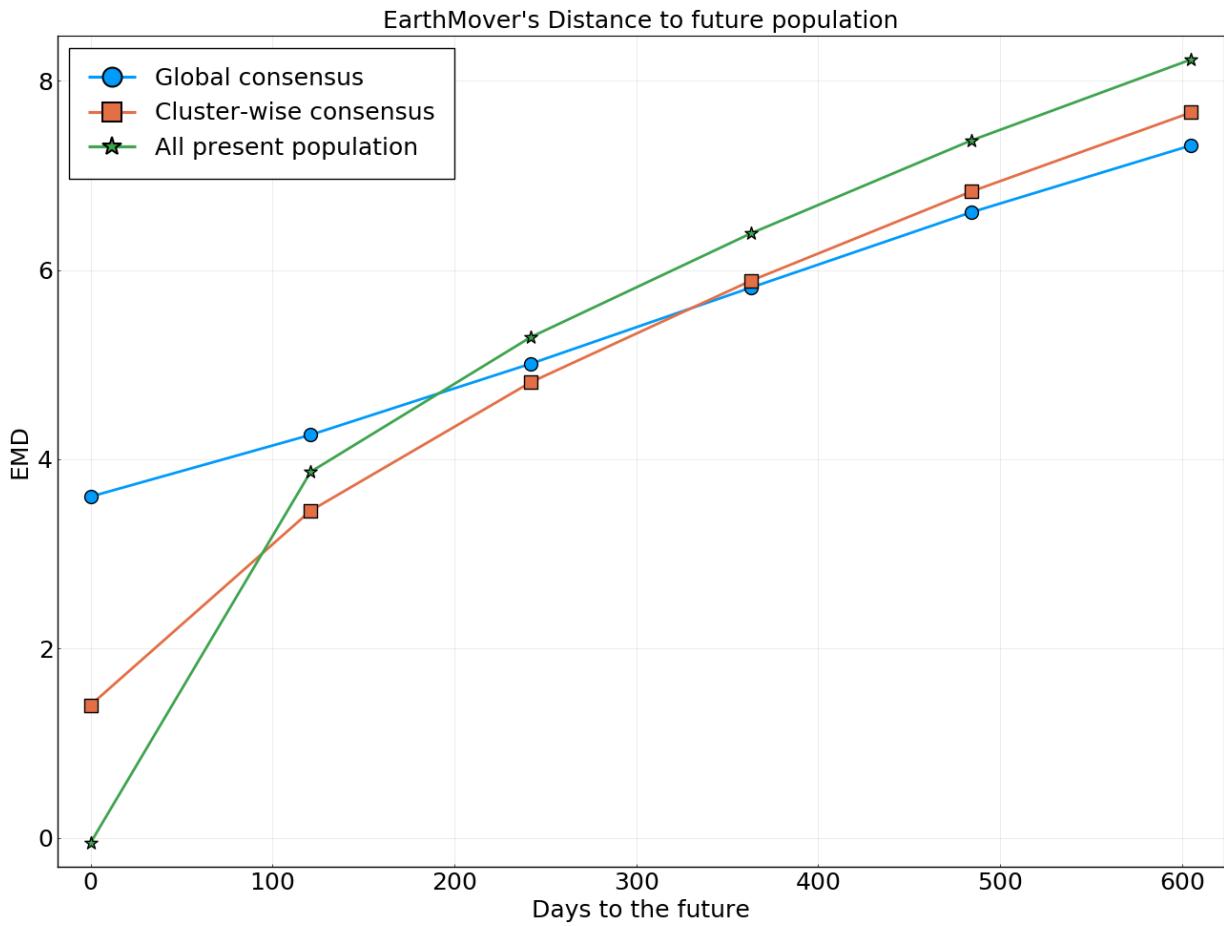


Figure S 19 Earth mover's distance to the future population for different predictors. A present population consists of all A/H3N2 HA sequences sampled in a 4 months time window. Quantities are averaged over all possible "present" populations from the year 2002. Predictors are: **Global consensus**: Consensus sequence of the present population. Best long-term predictor for a structure-less neutrally evolving population. **All present population**: All sequences in the present population. Perfect predictor if the population does not change at all through time. **Cluster-wise consensus**: Consensus sequence for each cluster in the present population. Clusters are based on local maxima of the LBI. Sequences are assigned to a given cluster based on their tree branch-length distance to the corresponding local maximum.

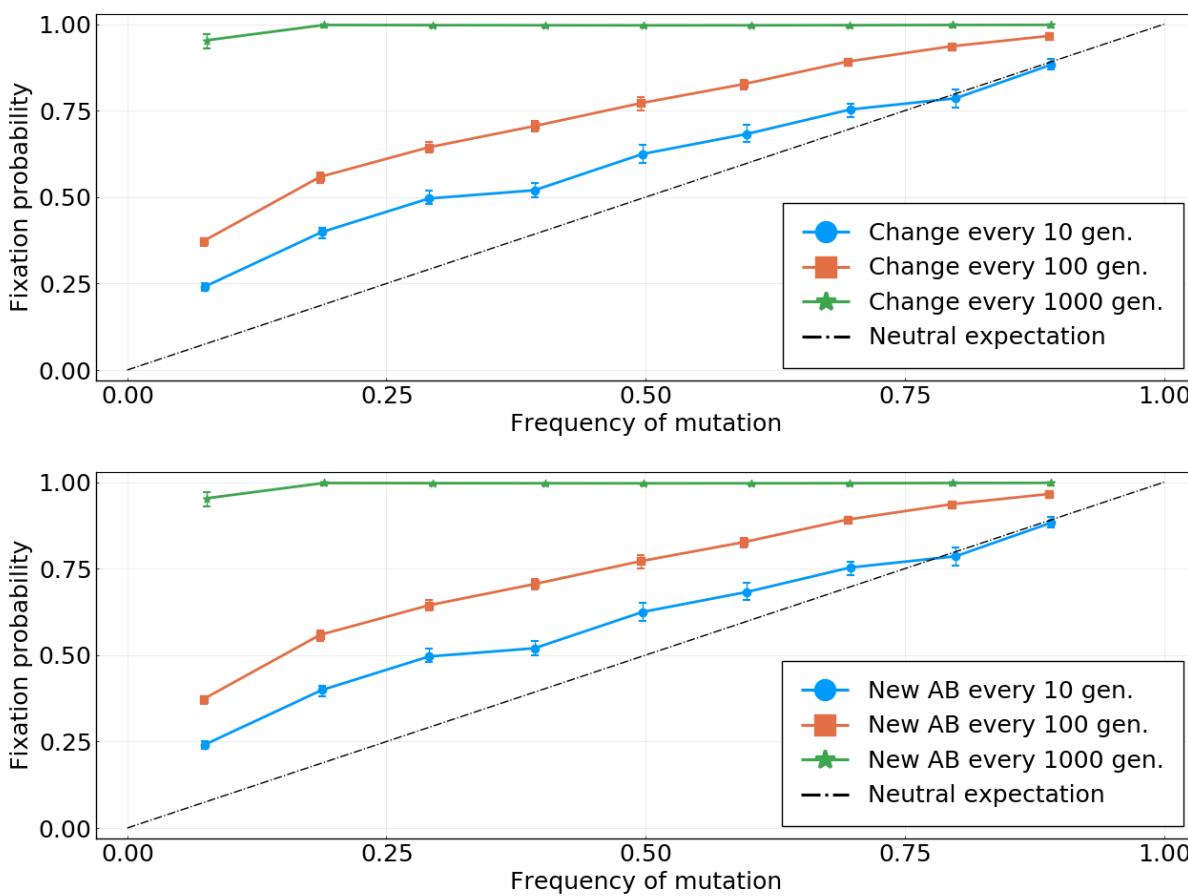


Figure S 20 Fixation probability as a function of frequency for the simulations discussed in the main text. **Top:** Simulation without antibodies. The three colored curves reflect different rate of change for the fitness landscape. Visual inspection of the frequency trajectories indicates a typical sweep time of  $\sim 400$  generations. **Bottom:** Simulation with antibodies. The different colored curves indicate the rate at which antibodies are introduced.

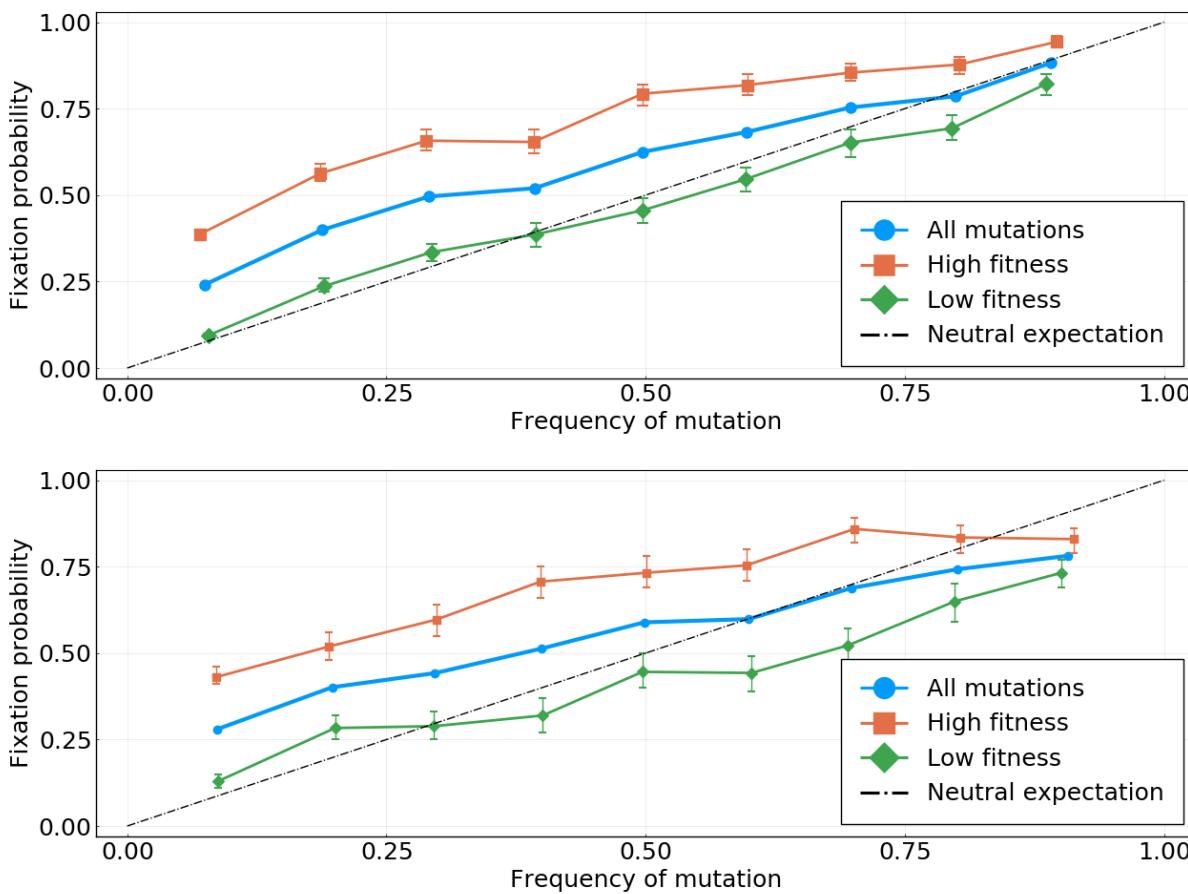


Figure S 21 Fixation probability as a function of frequency for the simulations discussed in the main text, with trajectories stratified according to real fitness values. “High” and “low” fitness classes are defined with respect to the median value. **Top:** Simulation without antibodies and with changes to the fitness landscape every  $dt = 10$  generations. **Bottom:** Simulation with antibodies, with a new antibody every  $dt = 10$  generations.