Short and long-read ultra-deep sequencing profiles emerging heterogeneity across five platform Escherichia coli strains

Citation for the original published paper (version of record):
Rugbjerg, P., Dyerberg, A., Quainoo, S. et al (2021). Short and long-read ultra-deep sequencing profiles emerging heterogeneity across five platform
Escherichia coli strains. Metabolic Engineering, 65: 197-206.
http://dx.doi.org/10.1016/j.ymben.2020.11.006

Metabolic Engineering 65 (2021) 197–206

Contents lists available at ScienceDirect

Metabolic Engineering

journal homepage: www.elsevier.com/locate/meteng

Short and long-read ultra-deep sequencing profiles emerging heterogeneity across five platform *Escherichia coli* strains

Peter Rugbjerg [a,b,1], Anne Sofie Brask Dyerberg [a], Scott Quainoo [a], Christian Munck [b], Morten Otto Alexander Sommer [a,*]

[a] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Building 220, Kgs. Lyngby, Denmark
[b] Enduro Genetics ApS, Copenhagen, Denmark

ARTICLE INFO

Keywords:
Population dynamics
Bioprocess scale-up
Genetic heterogeneity
Nanopore sequencing

ABSTRACT

Reprogramming organisms for large-scale bioproduction counters their evolutionary objectives of fast growth and often leads to mutational collapse of the engineered production pathways during cultivation. Yet, the mutational susceptibility of academic and industrial *Escherichia coli* bioproduction host strains are poorly understood. In this study, we apply 2nd and 3rd generation deep sequencing to profile simultaneous modes of genetic heterogeneity that decimate engineered biosynthetic production in five popular *E. coli* hosts BL21(DE3), TOP10, MG1655, W, and W3110 producing 2,3-butanediol and mevalonic acid. Combining short-read and long-read sequencing, we detect strain and sequence-specific mutational modes including single nucleotide polymorphism, inversion, and mobile element transposition, as well as complex structural variations that disrupt the integrity of the engineered biosynthetic pathway. Our analysis suggests that organism engineers should avoid chassis strains hosting active insertion sequence (IS) subfamilies such as IS1 and IS10 present in popular *E. coli* TOP10. We also recommend monitoring for increased mutagenicity in the pathway transcription initiation regions and recombinogenic repeats. Together, short and long sequencing reads identified latent low-frequency mutation events such as a short detrimental inversion within a pathway gene, driven by 8-bp short inverted repeats. This demonstrates the power of combining ultra-deep DNA sequencing technologies to profile genetic heterogeneities of engineered constructs and explore the markedly different mutational landscapes of common *E. coli* host strains. The observed multitude of evolving variants underlines the usefulness of early mutational profiling for new synthetic pathways designed to sustain in organisms over long cultivation scales.

1. Introduction

Successful industrial biotechnological production requires scaling of cultures from laboratory flasks and benchtop fermentors to large bioreactors at volumes up to several hundred cubic meters (Nielsen and Keasling, 2016). In cultured populations of high-producing engineered organisms, the associated metabolic burden and inhibitions (production load) select for spontaneous production escape events in the load-carrying genes. This production load can be quantified as relative reduction in specific growth rate due to production and results in heterogeneous populations and production decline over time (Ikeda, 2003; Kwon et al., 2015; Rugbjerg et al., 2018a). Such genetic instability challenges bioprocess scale-up and requires specific and sometimes extensive solutions to yield robust production strains (Borkowski et al., 2016; Wehrs et al., 2019; Zelder and Hauer, 2000). However, little is known about the genetic escape paths of popular host strains: heterogeneity generally evolves in cultivations beyond bench-top scale and requires deep sequencing in order to massively profile in parallel. Elaborate proteome and metabolome profiling of such strains can guide the choice of host strain dependent on the metabolic requirements (Monk et al., 2016), as well as rapid screening of many candidate hosts variants (Kim et al., 2014). Yet, host-specific genetic scale-up problems are not readily evident after short cultivations because of the difficulties in connecting lab-stage strain development with the different conditions of late-stage large-scale production. This limits early quantitative assessment and early prevention of production declines in industrial bioproduction projects (Rugbjerg and Sommer, 2019).

Deep DNA sequencing is now emerging as a strategy to profile the

* Corresponding author.
E-mail address: msom@bio.dtu.dk (M.O.A. Sommer).
[1] Present address: Department of Biology and Biological Engineering, Industrial Biotechnology, Chalmers University of Technology, Gothenburg, Sweden

https://doi.org/10.1016/j.ymben.2020.11.006
Received 5 August 2020; Received in revised form 26 October 2020; Accepted 12 November 2020
Available online 24 November 2020

development of genetic heterogeneity early in the strain design process. This allows timely prevention of potential genetic design flaws by redesigning gene and expression constructs before being confined to the physiological complexity of highly developed late-stage strains. A population of metabolically burdened cells generally declines in production because of spontaneous escape mutations in critical pathway genes and subsequent positive selection of the escapees (Rugbjerg et al., 2018a). Designing robust and scalable bioproduction strains requires significant engineering work and benefits from several strategies including adaptive laboratory evolution (Sandberg et al., 2019), synthetic addictions to products (Rugbjerg et al., 2018b), dynamically-regulated pathways and prevention of mutations in vulnerable genetic sequence architectures (Bull and Barrick, 2017; Ceroni et al., 2018; Xu, 2018). Despite this, the specific susceptibility of common host strains to heterogeneity has not been investigated and strategies are lacking for efficient profiling of long-term genetic stability of engineered bioproduction strains.

We previously applied short-read deep DNA sequencing to a 90-generation constantly-growing time-course of cultured production populations and found that a major gene escape mode in *Escherichia coli* TOP10 is gene disruption by diverse mobile insertion sequence (IS) elements (Rugbjerg et al., 2018a). Specific IS subfamilies such as IS10 and IS186 more frequently cause gene disruption than others, with some target site preference (Fan et al., 2019). This prompted us to ask whether short and long-read ultra-deep DNA-sequencing could be employed to profile and screen for strain effects resulting from differences in DNA repair systems and IS element abundance across host strains.

Single nucleotide polymorphisms (SNPs) and IS element transposition are detectable at subpercentage frequencies using short-read population sequencing (Rugbjerg and Sommer, 2019). In contrast, detecting complex structural variations such as inversions, duplications and deletions is challenging. However, these complex mutations may play significant roles in the evolutionary response to an engineered production pathway (Deatherage et al., 2015). Recent technical and computational advances in nanopore long-read sequencing now allow us to detect and analyze for these variants (Sedlazeck et al., 2018b).

In industry and academia, popular *E. coli* chassis strains include K-12 MG1655, which is the most well-characterized type strain. K-12 W3110 is genetically and metabolically similar yet differs in key metabolic reactions and is more frequently used industrially (Hayashi et al., 2006; Monk et al., 2016). TOP10 is a widely used K-12 cloning host due to high transformation efficiency and deficiency of *endA* and *recA*. TOP10 and its close relative DH10B are often used in academic metabolite production and synthetic biology studies (Ceroni et al., 2018; Martin et al., 2003). However, they host a high number of IS elements as seen in the genome sequence (Durfee et al., 2008), potentially limiting the applications beyond standalone cloning. BL21(DE3) is a popular B strain for high-level protein expression (Jeong et al., 2009), while strain W is fast growing and can utilize sucrose (Park et al., 2011).

In this study, we experimentally simulated large-scale cultivations with these five common *E. coli* host strains by serial passages of growing cultures to avoid stationary phase transitions. We combined long-read and short-read sequencing platforms to compare mutational modes for two heterologous metabolic pathways to mevalonic acid and 2,3-butanediol, respectively. Mevalonic acid is a precursor to isoprenoid fragrances, plastics and medicine. It can be synthesized in *E. coli* from acetyl-CoA by overexpressing native acetyl-CoA acetyltransferase and two heterologous enzymes (Xiong et al., 2014). The potential biofuel 2,3-butanediol has low microbial toxicity and can be synthesized in *E. coli* from pyruvate by overexpression of three heterologous enzymes (Xu et al., 2014).

2. Results

2.1. Different chromosomal IS compositions and production loads in common E. coli production strains

We compared the modes by which genetic heterogeneity evolved in five common metabolite production hosts in an experimental simulation of large-scale, long-term (90 cell generations) bioproduction (Fig. 1A). We first conducted a short-term phenotypic comparison after direct cultivation from single cell (approximately 20 generations). The five *E. coli* strains, BL21(DE3), MG1655, TOP10, W, and W3110, produced the two case-products, 2,3-butanediol and mevalonic acid, at different titers of 0.5–1.1 g/L and 0.2–0.5 g/L respectively (Fig. 1B), and with different degrees of production loads (7–30%), defined as the percentwise reduction in specific growth rate compared to the respective non-producer wildtype (Fig. 1B). The degree of production load positively correlated with production titers, though not at perfect linearity with R^2 of 0.66 and 0.48 respectively (Fig. 1B).

Using the available genome sequences for the host strains and literature, we quantified IS copy numbers (Fig. 1C) (Methods). We found large variations in the distributions of the different IS element subtypes in the five strains potentially translating into different rates of IS-based production decline. Some strains such as TOP10 and BL21(DE3) harbored up to 66 copies and many different subtypes of IS elements (Fig. 1C). In contrast, W harbored a total of only 16 IS elements (Fig. 1C). We also screened the reference IS compositions of two other popular academic *E. coli* strains, Crooks and DH5-α, and found IS compositions to be well represented by the selected five strains, with DH5-α hosting both IS1 and IS10 (Fig. 1C) previously implicated in high transposition rates (Rugbjerg et al., 2018a; Sousa et al., 2013).

2.2. Host-specific genetic declines in long-term engineered mevalonic acid production

Next, to simulate large-scale industrial growth for 90 generations, we serially passaged mevalonic acid-producing cells strictly in growing phase every 12 h to avoid selecting for industrially rare stationary-phase culture transitions and resulting lag phases. This scheme also kept the growing cultures synchronized, similar to our previous experimental simulations (Rugbjerg et al., 2018a). Recent experimental simulations in yeast have incorporated serial passaging regimes to simulate long-term cultivation (D'Ambrosio et al., 2020; Lv et al., 2020). The basic setup is attractive by resembling industrial stability tests and is likely a fair approximation of the actual scale-up, though it does not capture physical changes due to the large-scale operation, such as oxygen and carbon gradients and starvation (Takors, 2012) (Wehrs et al., 2019).

Over the large-scale mevalonate production simulation, *E. coli* TOP10 emerged as the fastest-declining production strain (Fig. 2, and Supplementary Fig. S1). Among four parallel populations of each host, the TOP10 populations fully lost the mevalonate-producing phenotype within 75 generations of culture corroborating our previous results (Fig. 2) (Rugbjerg et al., 2018a). The decline in mevalonate production was accompanied by rising population growth rates (Fig. 2). Mevalonate-producing BL21(DE3) declined slower than TOP10 (Fig. 2) despite a higher measured production load (Fig. 1B). This result indicated a slower spontaneous escape rate for BL21(DE3). In closely related MG1655 and W3110, production declines of 20% from the starting titer were seen at the last sampling point of 90 generations (Fig. 2). *E. coli* W did not show declines in production but also had a modest production load and titer (Fig. 1B).

2.3. 2,3-Butanediol production declines to an intermediate-level production state

We engineered the five *E. coli* host strains to produce 2,3-butanediol by transformation with the previously developed pET-RABC plasmid (Xu

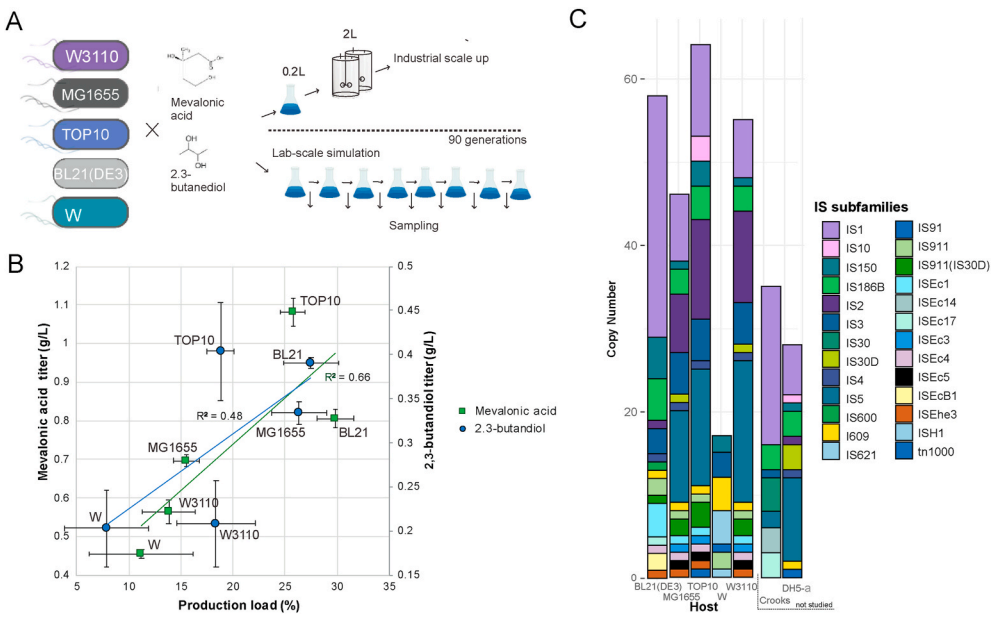

Fig. 1. Concept of the study exploring stability and heterogeneity profiles of five different, commonly used academic and industrial *E. coli* strains. A) Industrial long-term cultivation of strains was experimentally simulated by serial passaging of growing cultures and subjected to time-lapse sampling to study population dynamics when producing mevalonic acid or 2,3-butanediol. B) Prior to long-term cultivation, both product pathways confer a production load (percentwise growth rate reduction to wildtype) that is expected to amplify the individual rates of escape and generally scales with initial production titers (error bars depict s.e.m., n = 3). C) Common *E. coli* host strains differ in diversity and chromosomal copy number of IS elements. IS elements reported for the reference genomes of the five host stains experimentally studied and compared to two other popular host strains Crooks and DH5-α.

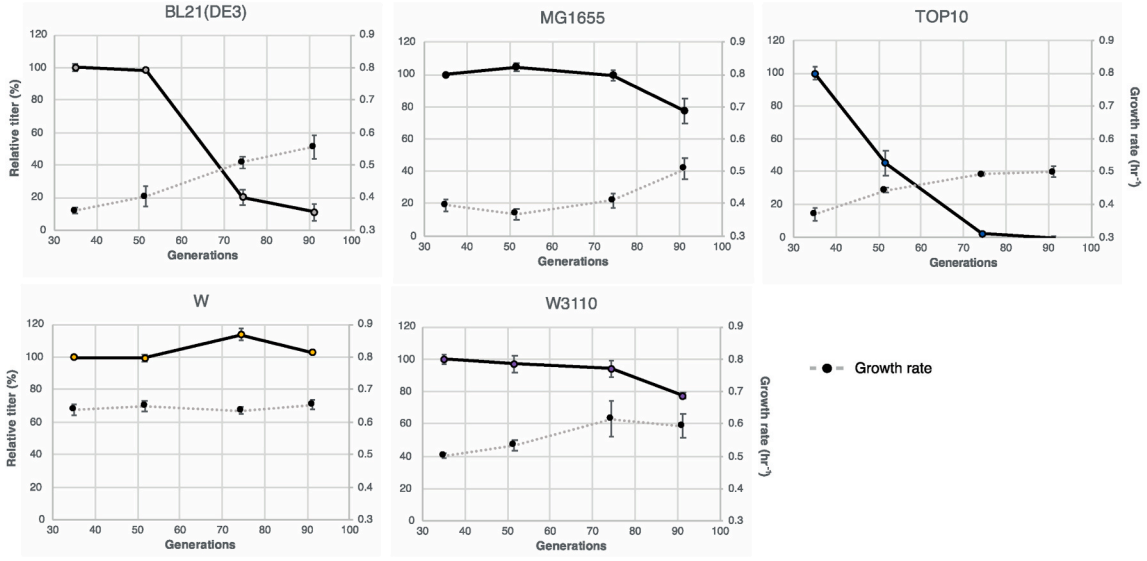

Fig. 2. Declines in mevalonate production titer (solid line) in experimentally simulated large-scale production differ in five *E. coli* platform strains and correlate with rising growth rate (dashed line). For all sample points, mean is shown with error bars representing standard error of the mean (n = 4).

et al., 2014). pET-RABC introduces the three heterologous pathway genes *budA, budB, budC* in an operon under a constitutive promoter along with non-utilized remnant genes *lacI* and *lysR*. We used the same experimental setup applied for mevalonic acid to simulate large-scale production using constantly growing, serially passaged cultures producing 2,3-butanediol for approx. 90 generations (Methods).

In all strains, 2,3-butanediol production declined, reaching intermediate levels at around 60% of the starting titer after 90 generations (Fig. 3 and Supplementary Fig. S2). TOP10, MG1655 and W3110 cultures went through a dynamic period at generations 50–75 followed by a more stable plateau. These adaptations were accompanied by an increase in specific growth rates (Fig. 3). In contrast, no significant increase in fitness accompanied the production declines of BL21(DE3) and W. These different dynamics over the large-scale simulation may also be somewhat reflected by the relatively large differences in initial production titer of the five strains (Fig. 1A and Supplementary Fig. S2), for example we measured the lowest initial production titer and load of 2,3-butanediol and mevalonic acid in W, but also the highest stability.

2.4. Transcription initiation region marks a strong SNP and IS insertion hot spot in mevalonate-producing TOP10 and BL21(DE3)

To compare differences in genetic heterogeneity among the five host strains, we first ultradeep-sequenced the mevalonate production plasmid (pMVA1) populations at the beginning and end of the experimentally simulated large-scale fermentation. For each host and timepoint, three of the four parallel-cultivated populations were sequenced using short paired-end reads (2x150 base pair) with a minimum sequencing coverage of 4,000x (Methods). Matching our phenotypic observations of mevalonic acid production decline, the pathway genes especially in TOP10 and BL21(DE3) contained a high degree of mutation after 81 generations of cultivation. The mevalonic acid production plasmid pMVA1 we used overexpresses the three pathway genes *atoB, mvaS* and *mvaE* in an operon using the constitutive "Anderson" J23100

promoter. At the 81-generation time point, approx. 10% of the pathway population contained SNPs (including short deletions) close to the transcription initiation region of production pathway genes (Fig. 4A). This mutational hot spot contained several SNPs (Fig. 4A), mainly loss-of-function. However, over the course of the 81 generations, particular SNPs and indels enriched more often across replicates and host strains (Supplementary Fig. S3). This trend indicates a higher spontaneous mutation rate for certain SNPs (e.g. short homopolymers) in addition to the mutation hot spot around the transcription initiation region, as previously found in other biological systems (Jinks-Robertson and Bhagwat, 2014). In the most mutated host strains, BL21(DE3) and TOP10, specific pathway SNPs were found to enrich in all three sequenced replicates (Supplementary Fig. S3). By sequencing populations at two time points, we could confidently identify SNPs that predominantly or uniquely enriched in specific strains (Supplementary Fig. S3). Time resolution also improved the ability to distinguish true and recurring false SNPs at such low frequency (0.15–1%), since the latter are strain-agnostic sequencing artefacts and remain at constant frequency over time (Supplementary Fig. S3). At a 0.15% SNP detection limit, none of the reported early SNPs rose to higher frequencies following the large-scale cultivation simulation, which indicated that dominant late-stage SNPs could not be detected through early culture sequencing with a 0.15% detection limit (Supplementary Fig. S3), which however is also not particularly deep from the perspective of predicting microbial evolution early on. In fastest-escaping TOP10 populations, most enriched SNPs were in the short constitutive J23100 promoter (2.8% ±1.7 of population) or *atoB* (7.8 ± 2.1% of population), the first gene in the operon of the engineered biosynthetic pathway (Fig. 4A and Supplementary Fig. S3). This result indicated selection for these mutations leading to complete loss of engineered mevalonic acid production. Similarly, we detected a number of IS insertions by analyzing split-end reads (Methods). IS insertions were also targeted around the transcription initiation region and largely impacted the J23100 promoter and *atoB* gene first in the biosynthetic operon (Fig. 4C), reaching respectively 7 ± 2% and 12 ± 3% of the populations. The heterologous *mvaS* and *mvaE* genes of the operon, encoding HMG-CoA synthase and reductase, were largely intact (IS disruptions in respectively 6 ± 6 and 0% of the populations).

The p15A origin also marked a host-specific mutation hot spot with substantial enrichment of SNPs in W3110 and W replicates to approx. 40% of the population (Fig. 4A), thus showing that the approx. 50 mutations largely co-occurred. Notably, the same mutational pattern could be observed in BL21(DE3) replicates (Fig. 4A) in the 0.1–1% population frequency. It is tempting to speculate that these trends drive a change in plasmid copy number and. An interaction with native plasmids, e.g. the 100 kb large pRK1 in W, is less likely, since the pattern was also seen in other strains. However, these hypotheses were not tested.

2.5. Limited genetic heterogeneity behind intermediate 2,3-butanediol production decline

To investigate if the observed phenotypic declines in 2,3-butanediol production was due to genetic heterogeneity, we ultradeep-sequenced the pET-RABC production-plasmid populations from three of four parallel lineages from the five host strains before and after the experimentally simulated long-term fermentation using short paired-end reads (2x150 base pair). Only a limited number of SNPs accumulated to more than 1% population frequencies in the pathway genes, affecting TOP10 and W over 81 generations (Fig. 4B). However, many SNPs reached subpercentage frequency even when disregarding presumably artificial SNPs also detected at the early sequencing time-point (25 generations) (Methods) (Fig. 4B) (Supplementary Fig. S4). In contrast, higher frequencies of SNPs in both specific positions and specific host strains enriched over the time course for the mevalonate producing strains (Supplementary Fig. S3). The acetolactate synthase gene *budB* was the most mutated pathway gene even though acetolactate toxicity is not expected in *E. coli* (Aristidou et al., 1994). Notably, IS elements were not detected in pET-RABC plasmid populations from any replicates of the tested five hosts using short-read sequence data (Fig. 4D), but only subsequently two <1% IS insertions were detected in long-read data for TOP10. This could indicate that mutation formed at lower rates in the pET-RABC system since the measured production load was approximately the same for the two case pathways.

Interestingly, also the pMB1 origin of replication in pET-RABC were marked as mutation hot spots for enrichment of SNPs in the same two host organisms W and W3110 (Fig. 4B). Even if the p15A and pMB1 origins are mechanistically very different, such changes could e.g. drive copy number differences over time. Finally, it is possible that chromosomal changes could enrich over time, e.g. as in yeast (D'Ambrosio et al., 2020), though not seen in a previous study of mevalonic acid production decline (Rugbjerg et al., 2018a). To survey this at population-level, the

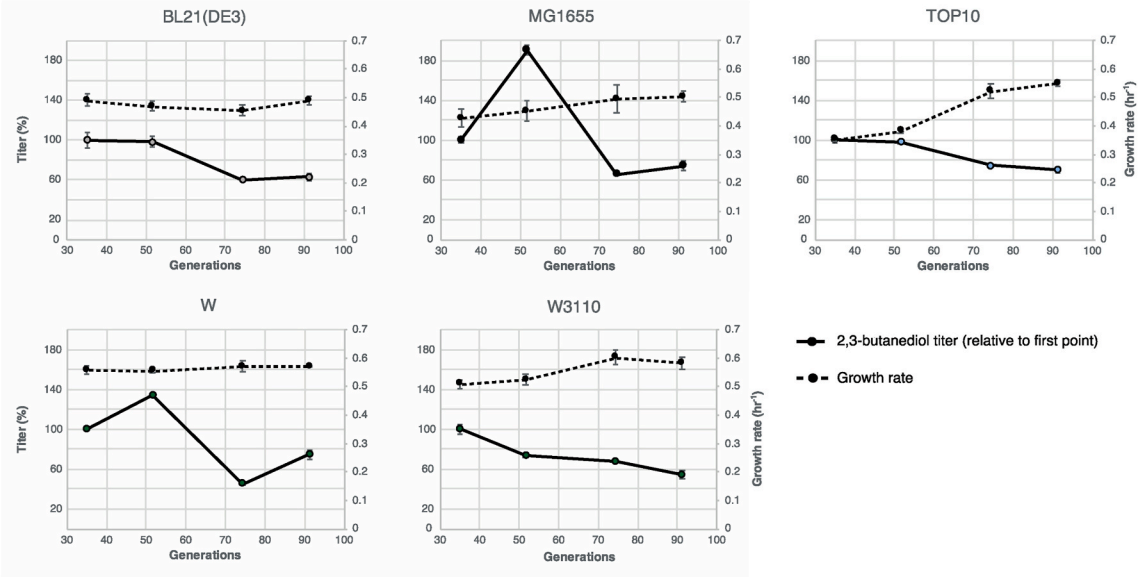

Fig. 3. Five *E. coli* platform strains engineered for 2,3-butanediol production show declining production titers in long-term production and rising growth rate (dashed line). For all sample points, mean is shown with error bars representing standard error of the mean (n = 4). Titers are relative to first measured point.

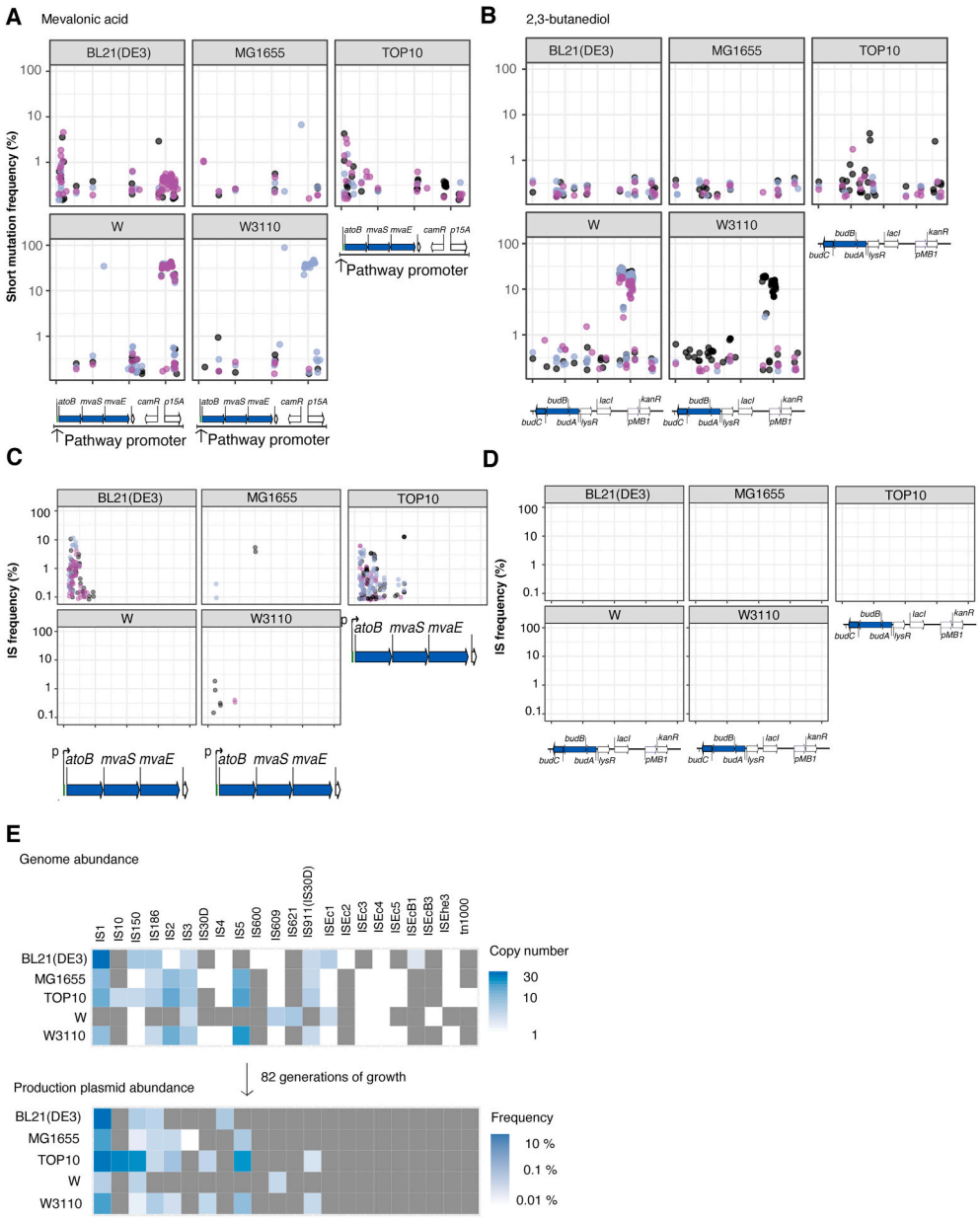

Fig. 4. Genetic heterogeneity enriched during long-term cultivation of five common *E. coli* host strains engineered to produce mevalonate and 2,3-butanediol respectively. Three parallel populations (indicated by color) were short-read ultra-deep sequenced following respectively 82 and 81 generations of cultivation without stationary phase. A) Population frequencies of SNPs and short deletions in mevalonic acid producing cells. B) Population frequencies of SNPs and short deletions in 2,3-butanediol-producing cells. C) Position-resolved population frequencies of IS elements indicated by broken sequencing reads mapping to the reference production construct and an IS element. One IS insertion normally produces two broken read points. D) No IS insertions were seen by short-read DNA sequencing of the 2,3-butanediol-producing populations. E) Heatmap shows presence of IS subfamilies respective host genomes and in mevalonic acid-plasmid populations following 82 generations of growth. Colors indicate values of individual biological replicates (n = 3). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

sequencing needed would be significantly higher to satisfy an average whole-genome coverage similar to the plasmid population of the present study, but it could bring forward important clues on the most pressured parts of native metabolism, on which the pathway drains substrate, energy and redox power.

2.6. *Long-read deep sequencing uncovers complex structural variants and a short inversion between naturally coded inverted repeats*

Short-read sequencing can accurately detect mobile element disruptions, SNPs and short insertions/deletions (indels) in deep-sequenced populations, however spurious split-end short reads in sequenced plasmid populations from all five host strains indicated that more structural variation might be taking place (Supplementary Figs. S5 and S6). We therefore explored if long-read sequencing could enable discovery of other structural variants within the production populations (Sedlazeck et al., 2018a). Specifically, we hypothesized that long reads permit the discovery of overlooked subpopulations carrying structural variation and correct for rare sequencing artefacts unique to short-read sequencing at subpercentage frequencies. We linearized purified plasmid populations from three lineages of each of the five mevalonic acid-producing host strains (Methods) and sequenced the populations using nanopore technology (Minion, Oxford Nanopore Technologies) for ultrahigh coverage (>10,000x) (Methods). We took advantage of the option for PCR-free library preparation to benefit from its intrinsic independence of DNA polymerase and avoid polymerase-related artefacts. Such artefacts may extend beyond SNPs and remain a major disadvantage of sequencing-by-synthesis short-read platforms. We reference-mapped the long reads using the long-read mapper NGMLR and detected structural variation using Sniffles (version 1.0.7) (Sedlazeck et al., 2018b) (Methods).

In agreement with the observed phenotypic dynamics for mevalonic acid production titer and growth rate (Fig. 2), structural variations were most highly enriched in the TOP10 populations but were also significantly frequent in BL21(DE3), MG1655 and W3110 and to some extent in W (Fig. 5A). Consistent with our short-read data, nanopore sequencing detected many insertions with IS elements into *atoB* in the three sequenced TOP10 replicate lineages (9 ISs above 1% using long

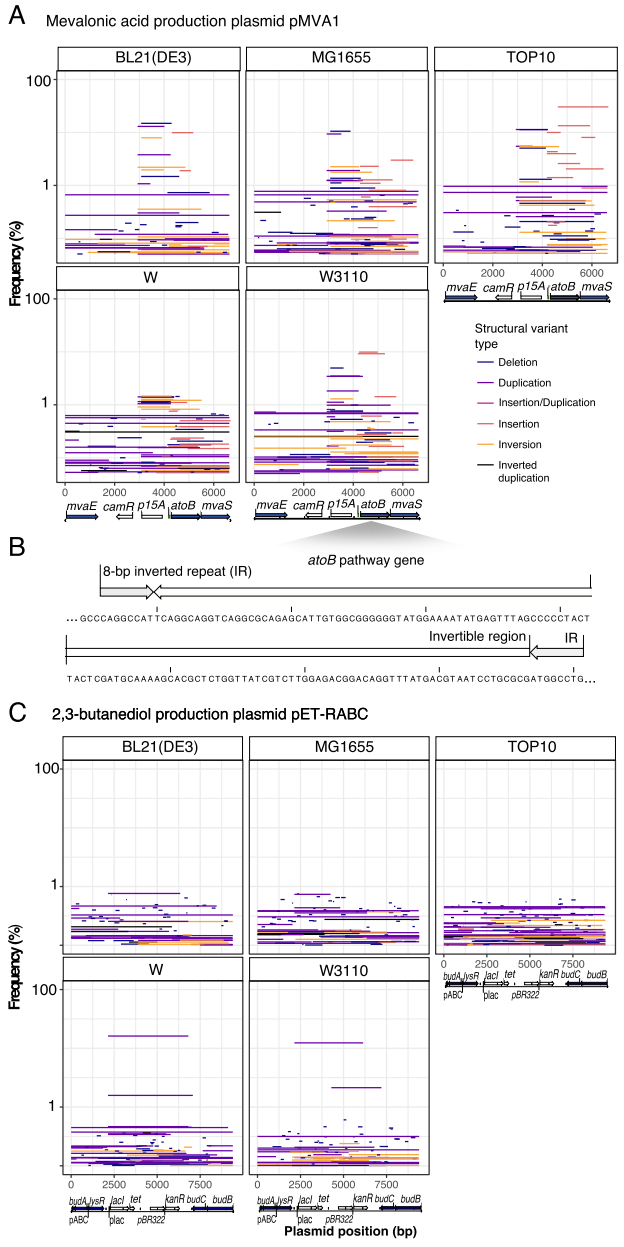

Fig. 5. Ultra-deep long-read nanopore DNA sequencing reveals structural variants in production plasmids of populations of five *E. coli* chassis strains. Populations producing mevalonate and 2,3-butanediol were analyzed after respectively 80 and 81 generations of un-interrupted growth to simulate large-scale production. A) Structural variation detected in mevalonic acid production populations. Data from three sequenced biological replicates for each host strain show different modes of complex structural variation. Each line was adjusted according to its best left-sided split end (Sniffles prediction). For insertions, line length indicates insertion length. Colors indicate respective structural variant forms (Methods). B) Schematic of inversion within pathway gene driven by short inverted repeats (IRs) naturally coded in the mevalonic acid pathway gene *atoB* leading to prematurely terminated pathway enzyme variants in subpopulations of *E. coli* W3110. C) Structural variation detected in 2,3-butanediol production populations. Data from three sequenced biological replicates for each host strain show different modes of complex structural variation. Each line was adjusted according to its best left-sided split end (Sniffles prediction). For insertions, line length indicates insertion length. For TOP10, rare IS1 insertion emphasized with bolder strength. Colors indicate respective structural variant forms detected in the populations of a randomly selected replicate (Methods). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

reads, 10 ISs above 1% using short reads).

On the other hand, Sniffles predicted structural variants not seen when inspecting split-end short reads (Supplementary Fig. S5 and Fig. S6), e.g. near the plasmid origin of replication (Fig. 5). There, co-developing complex structural variations of deletion, duplication, or inversion events were enriched and reached frequencies of approx. 20% in TOP10 and BL21(DE3) replicates.

At end of the experimentally simulated long-term fermentation, we found that approx. 0.2% subpopulations in two of three long-read-sequenced *E. coli* W3110 lineages carried a 134-bp inversion confined by 8-base pair short IRs naturally present in the coding sequence of pathway gene *atoB* (Supplementary Text). This inversion introduces two premature stop codons that disrupt mevalonic acid production (Fig. 5B). Knowing the specific position of the inversion, we also searched for it in the split-end short-read data and found it uniquely in data from the same two W3110 lineages, thus cross-validating its existence. Since the short-read length was constrained to 150 base pairs, no single short read encompassed the full IR-inversion-IR, supporting why it had been difficult to predict *de novo* using our 150 bp paired-end reads. *de novo* discovery of the inversion may therefore have been aided by longer short reads. Further, inspection of paired short reads identified pairs covering the inversion, which confirmed that the inversion was present in the pathway population of W3110 lineages. While inversion as sequencing artefact has not previously been reported on the short-read Illumina platform, IR-related secondary structures do promote certain erroneous SNPs on this platform (Nakamura et al., 2011).

Long-read sequencing predicted the existence of several complex forms of structural variations such as co-appearing duplications and deletions around the p15A origin of replication in pMVA1 (Fig. 5A). These structural variants of the origin of replication did not limit to W and W3110 as we found for the SNP hot spots we identified using short reads, thereby indicating a different mechanism. Since an origin is essential to replicate a plasmid, this predictions of deletions supports the observations that plasmid copies can reside in multimers that undergo complex recombination (Chang and Cohen, 1978). Such tandem duplicated plasmid is not directly detectable after necessary linearization for nanopore sequencing though we also see *in vitro* <1% spontaneous full-length duplications at the linearization points (Fig. 5A). The existence of intra-plasmid recombination could only partially be described using split-end short reads (Supplementary Fig. S5). Some predicted structural variation in the pMVA1 plasmids seemed to interact with the boundary of a 144-bp incomplete, undescribed remnant of IS1 that we found close to the p15A origin dating back to its initial domestication in pACYC184 (Chang and Cohen, 1978). We find this IS1 remnant is present in many but not all p15A-based plasmids deposited in Genbank, indicating non-essentiality to the origin. It is uncertain whether this IS remnant plays a detrimental role to stability, but this stretch can presumably recombine with any intruding second IS1. As use of long-read nanopore sequencing grows, we will potentially gain more information about potential errors of this method to help interpret read data. We also detected minor tandem duplications and inversions (Fig. 5A) along with many non-enriched small deletions that may have been nanopore-specific sequencing artefacts (Sedlazeck et al., 2018a). We also sequenced the end-point (81-generation) 2,3-butanediol production plasmids (pET-RABC) with ultra-deep nanopore long reads to see if undiscovered mutation modes could be resolved by this technology for this pathway (Fig. 5C) (Methods). In two TOP10 populations, nanopore reads allowed for identification of two very rare (0.02% and 0.2%) instances of IS1 insertion into different positions in *budA*, the first gene of the pathway operon that encodes α-acetolactate decarboxylase (Fig. 5C). These rare ISs were overlooked by the short-read sequencing, but otherwise no IS disruption was detected in pET-RABC. In addition, we detected a 4.7 kb long backbone duplication, effectively doubling *lacI*, *kanR* and the pBR322 origin of replication (Fig. 5C), which was unique to long-read analysis and conserved in all 15 sequenced lineages of the five host strains. In one lineage of W and W3110, high frequencies of the

duplication at 12% and 16% were seen, whereas the remaining cases varied between 0.1 and 2.2%. Such intra-plasmid recombination was only partially detected by the corresponding split-end short reads (Supplementary Fig. S5).

3. Discussion

In this study, we systematically compared the emergence of genetic heterogeneity within five *E. coli* host strains that are commonly used in academic and industrial bioproduction. In an experimentally simulated large-scale production of two case products, we found that these hosts differ substantially in their mutational modes and frequencies. Specifically, we saw certain host strains were more susceptible to decline driven by differences in activity by IS elements with IS1 and IS10 transposition being particularly active in agreement with previous studies in MG1655:tn10 and TOP10 (Fan et al., 2019; Rugbjerg et al., 2018a; Sousa et al., 2013). However our data also indicates that the basis for the rate of decline is product/plasmid-specific, which may be associated to the mechanism of plasmid segregation (discussed more below). We also found that differences in production load caused by the same metabolic pathway in different host strains largely varied with production titer, thereby underpinning the engineering tradeoff that higher load can lead to higher instability. Based on these results, we propose that TOP10 is a less stable metabolite production host because it contains more active IS subfamilies IS10 and IS1. Notably, several other popular *E. coli* strains also harbor active IS10 such as DH5-α and Nissle 1917 (Fan et al., 2019). However, the vehicle of gene propagation also appears to matter, given the higher stability of the pET-RABC based production. As has been discussed previously, the imbalanced segregation of plasmid copies at cell division could elevate the spread of mutation (Rugbjerg and Olsson, 2020; Tyo et al., 2009). In this regards, chromosomal integration may provide better stability over time since every gene copy is mutationally independent. Recent studies have presented CRISPR/Cas9-aided mutation or silencing strategies targeting the transposase genes driving IS transposition (Geng et al., 2019; Nyerges et al., 2019; Umenhoffer et al., 2017). Based on our results, abolishing only subsets of the typical 20–60 IS copies (e.g. IS1 and IS10 subfamilies) may be sufficient in many typical bacterial production organisms. In particular, IS silencing appears interesting for strain diagnostics, but may be unrealistic in an actual production environment due to the typical burden associated with CRISPR/Cas9 expression. Further, using DNA deep-seq data, we find a considerable higher mutational activity (ISs and SNPs) around the transcriptional initiation region (Fig. 4C). This could be due to a selective advantage of abolishing even the first gene of the pathway operon (*atoB*) along with the downstream coded genes, yet physiological studies have clearly linked the toxicity of mevalonate production to the expression of the second operon gene, the heterologous HMG-CoA synthase (coded by *mvaS*) (Kizer et al., 2008). This suggests that promoter areas may be spontaneously more mutagenic due to the high transcriptional activity (Jinks-Robertson and Bhagwat, 2014), and these could be a potential target for strain engineers looking to scale processes.

These results demonstrate the significant advantage of profiling the genetic stability of individual pathway designs using ultra-deep DNA sequencing prior to long-term usage. Compared to sequencing isolates using e.g. Sanger sequencing, deep sequencing provides a fuller picture of frequently competing mutational modes. Specifically, we find that a combination of short-read and long-read sequencing allows for a broader capture of mutational modes.

2,3-butanediol production was generally more stably maintained than mevalonic acid production. The production levels decreased towards a 40-percent lower level, indicating that this was sufficient for mitigating most of the burden associated with 2,3-butanediol production. Ultra-deep sequencing by short and long reads indicated that genetic heterogeneity had developed resulting in low levels of SNPs overall. Long-reads specifically predicted IS element insertions within TOP10 lineages and the formation of backbone duplications, which appeared in the same sequence in all strains but at variable frequencies (Fig. 5C), suggesting a specific but cryptic mechanism involving the p*lacI* genetic sequence, which was a remnant not directly utilized for pathway expression in pET-RABC.

Emphasizing the utility of hybrid ultra-deep sequencing, combining short-and long-read sequencing increased our ability to identify mutational modes at different frequencies and led to the discovery of a short but detrimental inversion within a particular host strain (*E. coli* W3110). While this low-frequency inversion rose to only <1% frequency over 81 generations, it indicates a type of heterogeneity that could be expected in strains when abolishing all IS transposition. It also highlights that even common 8-bp short IRs can promote inversions to generate frameshifts effectively truncating the *atoB* pathway gene. This switching is similar to reversible phase variation by promoter inversions that regulates antibiotic resistance and flagella synthesis in bacteria, driven by IRs up to several hundred base pairs long (Bi and Liu, 1996; Jiang et al., 2019). Inversions between IRs can be dependent on specific DNA invertases e.g. coded by genomic phages, as well as general homologous recombination driven by RecA (Darmon and Leach, 2014), which is present in all our studied host strains except TOP10. The detrimental inversion within the biosynthetic *atoB* gene suggests that IRs overrepresented in many coding genes across phylogenetic kingdoms pose a risk for engineered production pathway stability. In evolution of *S. cerevisiae,* split-end short read population sequencing has guided the confirmation of longer chromosomal inverted gene amplifications at short palindromic junctions as adaptive mechanism (Payen et al., 2014), underlining the recombinogenic risks of repeat structures. Similar to many sequenced genomes, we found an overrepresented number of short IRs (Cox and Mirkin, 1997; Lavi et al., 2018). We counted 17 short IRs less than 250 bp apart in the mevalonic acid pathway genes (SI Text). This pattern indicates conserved functional biological roles that may be sequence-context dependent (e.g., secondary structures around terminators and RNA switches prior to translation initiation sites) (Lavi et al., 2018). It remains unknown what makes W3110 prone to this inversion that abolishes the coded gene function. Interestingly, gene synthesis and codon optimization of pathway genes may yield the unintentional positive side effect of preventing inversion by removing short-spaced IRs present in the source sequences, as the chance of randomly introducing an IR is far lower than their elevated natural abundance (Cox and Mirkin, 1997; Lavi et al., 2018). The observation of within-gene inversion that disrupts pathway enzyme expression supports use of codon optimization in engineered gene constructs merely for the sake of removing secondary information pre-encoded into natural genes.

Volumetric scale-up is essential to most industrial bioproduction projects, and the evolutionary pressure on engineered genetic constructs means detecting sequence vulnerabilities early is important. In this study we leveraged two next-generation sequencing technologies to profile *E. coli* strain and sequence specific genetic heterogeneity that accumulate through prolonged cell generations equivalent to those experienced in large-scale production. We showed emergence of highly sequence and host-specific mutation especially in promoter and coding regions (Fig. 4 and Figs. S3 and S4) as well as host-specific pathway disruption by transposing IS elements. Ultra-deep sequencing of early cultures at 25 generations only identified low-frequency SNPs (0.15–4%) that did not rise in frequency, were present across host strains and replicated populations and likely constituted artificial mutations. Different higher-frequency SNPs were detected following 81–82 generations of uninterrupted cultivation indicating that those were true and that early-cultivation sequencing could not detect true SNPs (when applying the 0.15% detection level). Similarly, the observed recurrence of promoter and gene-specific SNPs (Supplementary Figs. S3 and S4) speaks for using deep-seq in the evaluation of expression construct designs.

Thus, both in terms of preventing SNPs and structural variation in production genes, our study encourages routine use of deep-DNA

sequencing combined with serial-passaging experiments for profiling stability and failure modes of different candidate host strains as well as expression sequence designs prior to scale-up to avoid recurring mutations. Other production organisms also harbor many IS elements (e.g. lactobacilli), and most production hosts would amenable to serial-passaging setups combined with deep sequencing to screen for these. Our results point to general engineering principles such as avoiding mutation-prone sequences including direct or inverted repeats and host strains harboring many highly active mobile elements. Finally, these data suggest substantial strain-to-strain variation in the long-term stability of production populations highlighting the importance of testing several different production hosts when developing a bioprocess.

4. Methods

4.1. Strains

The investigated strains were constructed by standard electroporation of the indicated plasmid (Table 1) into the following host strains originating from the indicated sources.

E. coli BL21(DE3) (Coli Genetic Stock Center, Yale).

E. coli MG1655 (Coli Genetic Stock Center, Yale), contained the thermosensitive plasmid pKD46, which we cured by cultivation at 37 deg. C overnight.

E. coli TOP10 (Thermo Scientific).

E. coli W (DSMZ).

E. coli W3110 (Coli Genetic Stock Center, Yale).

4.2. Plasmids

4.3. Media

For all cultivations, unless otherwise stated, standard 2xYT characterization medium was used: 10 g/L yeast extract (Sigma-Aldrich), 16 g/L tryptone (Bacto), 5 g/L NaCl (pH adjusted to 7.0) supplemented with either 50 μg/mL kanamycin (pET-RABC-carrying strains) or 30 μg/mL chloramphenicol (pMVA1-carrying strains).

4.3.1. Long-term cultivation by serial continuous growth of *E. coli* producing mevalonic acid or 2,3-butanediol

Four parallel 25-mL cultures were started in 50-mL aerated tubes from single colonies and grown at 30 °C with horizontal shaking at 250 rpm. After 12 h, 0.5 mL broth was transferred into 25 mL fresh medium and incubated under the same conditions for another 12 h. At each passage, the OD_{600} was recorded to determine the accumulated number of cell divisions (Supplementary Table S1) and 1.8 mL 50% glycerol stocks were stored at −80 °C and 1.8 mL culture was stored at −20 °C.

4.4. High-depth short-read (Illumina) DNA sequencing and analysis

Production plasmid populations were purified from each time point sample using a standard plasmid purification kit (Macherey-Nagel). Samples were prepared for sequencing using the Nextera XT v2 set A kit (Illumina) per manufacturer's instructions with the addition of two 'limited-cycle PCR' cycles. Sequencing was in pooled Miseq runs with 150-bp paired-end reading. CLC Genomics Workbench (version 8.5) was used for initial bioinformatics analysis. First, reads were mapped to the reference pMVA1 or pET-RABC sequence respectively. Broken aligned reads were identified using the CLC Genomics Workbench Breakpoint analysis tool to yield a table of consensus broken unaligned reads and their abundance (maximum three mismatches allowed in the mapped read region, p-value for the fraction of unaligned reads set to 0.0001) to obtain an initial overview of structural variation. SNPs and short deletions were called using the CLC Genomics Workbench Low Frequency Variant Detection tool with 1% required significance level and 0.25% minimum frequency. SNP frequencies in sequenced populations were calculated by division with their respective coverage values.

SNPs found in the plasmid backbone at >90% frequencies in the sample of the initial seed were regarded as present in the starting plasmid. Frequencies of IS elements in plasmid populations were found by the relative coverage upon mapping the reads to all potential IS elements and only regarding fully covered instances. Frequencies of IS-related broken reads were calculated as the number of reads split between IS and reference relative to the sum of perfectly and non-perfectly aligned reads. IS subfamily sequences and counts were collected from the respective strain genome sequencing studies with help of previous prediction (Kim et al., 2011) (GenBank accessions BL21(DE3): CP001509, MG1655: NC_000913, For TOP10, the identical or closely related DH10B: NZ_010473, W: CP002185, W3110: AP009048, Crooks: CP000946, For DH5-alpha, NEB5-alpha: CP017100).

4.5. High-depth long-read (Nanopore) DNA sequencing and analysis

From each culture passaging time point, 2-mL samples were re-cultured from frozen sample stocks under conditions identical to the evolution experiments. At least 30 ng/uL plasmid DNA in was extracted using a standard plasmid extraction kit (Macherey-Nagel) and linearized using the single-cutting BamHI enzyme. We utilized the BamHI site because it would linearize the reference plasmids only once, and is located between pathway coding sequences, thus lowering risk for being involved in mutation. Briefly, multiplexed library preparation for nanopore sequencing was adapted from previous protocols (Quick et al., 2017) and thus end-prepared plasmid DNA was barcode- and adapter-ligated (Supplementary Text). A sequencing flow cell (R9.4.1; FLO-MIN106.1, Oxford Nanopore Technologies) was primed and loaded with the diluted library according to the manufacturer's instructions. Sequencing was performed with live base calling and stopped after three days or when the number of actively sequencing pores containing a single strand fell below 10. Resulting fastq files were aligned to the reference plasmid sequence using NGMLR (version 0.2.7) and analyzed for structural variation using Sniffles (version 1.0.7) (Sedlazeck et al., 2018b). Direct Sniffles output is available as Supplementary files. The sequencing data will be available via the ArrayExpress repository.

4.5.1. Measurement of 2,3-butanediol and mevalonic acid production by HPLC

At each culture passage, 900 μL medium was mixed with 900 μL 50% glycerol and stored at −80 °C. Following the simulated fermentation, each population sample from a 25-μL glycerol stock was used to inoculate 15 mL medium for incubation at 30 °C with shaking at 250 rpm for 54–58 h. Following incubation, 300-μL aliquots were treated with 23 μL 20% sulfuric acid. Samples were vigorously shaken and centrifuged at 13,000 g for 2 min. Supernatant (medium) samples were injected into an Ultimate 3000 HPLC running a 5 mM sulfuric acid mobile phase (0.6 mL/min) on an Aminex HPX-87H ion exclusion column (300 mm × 7.8 mm, Bio-Rad Laboratories) at 50 °C. A refractive index detector was used for detection. Standard curves for 2,3-butanediol and mevalonic acid were generated with respectively 2,3-butanediol and mevalonolactone (Sigma-Aldrich) dissolved in 2xYT medium supernatant from an engineered nonproducing strain incubated under same conditions.

Table 1
List of plasmids.

Plasmid	Features	Metabolite end product	Reference
pET-RABC	p_{con}-*budABC*, *lysR* *kanR*, pBR322	2,3-butanediol	Xu et al. (2014)
pMVA1	p_{J23100}-*atoB-mvaS-mvaE*, *camR*, p15A	Mevalonic acid	Rugbjerg et al. (2018b)

4.5.2. Measurement of growth rates and calculation of production load

To measure growth rates, 1.5-μL aliquots of stationary-phase cultures e.g. grown for productivity analysis (as described above) were used to inoculate 200 μL medium in microtiter plate wells. Plates were sealed with Breathe-Easy polyurethane seals (USA Scientific) and incubated with "fast" continuous shaking in an ELx808 kinetic plate reader (BioTek), which measured the OD_{630} value every 10 min.

Background-subtracted OD_{630} values were computed using measurements from uninoculated wells. Local growth rates were computed for each background-subtracted OD_{630} value by regression in rolling windows of five measurement points and background-subtracted OD_{630} values (Rugbjerg et al., 2018a). To represent growth rates, the third maximum local growth rate was reported. Production load was calculated as the percentwise reduction in growth rate compared to the same host strain grown without the respective production plasmid.

4.6. Replicates

Throughout the study, biological replicates were used for calculation of mean and s.e.m. In serial-passaging experiments, the biological replicates refer to independent lineages cultivated in parallel from different ancestral single colonies.

Data availability

Deep-sequencing data from the study has been deposited in ArrayExpress (accession no. E-MTAB-9800).

SNPs predicted from Illumina data (.csv) as supplementary material.

Broken read signatures predicted from Illumina data (.csv) as supplementary material.

Author contributions

All authors designed research, A.D. and P.R. performed research except for nanopore sequencing by S.Q. All authors analyzed data and wrote the paper.

Declaration of competing interest

C.M. and P.R. hold financial interests in Enduro Genetics ApS. All other authors declare no competing interest.

Acknowledgements

We thank Rebecca Lennen for valuable suggestions, Alexandra Hoffmeyer for skilled handling of Illumina sequencing and Cuiqing Ma (Shandong University) for kindly providing pET-RABC. Funding: The research leading to these results has received funding from the Novo Nordisk Foundation, Denmark, grant number NNF10CC1016517. The funding source had no role in design of the study, collection, analysis and interpretation of data nor in the decision to submit the article for publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymben.2020.11.006.

References

Aristidou, A.A., San, K.-Y., Bennett, G.N., 1994. Modification of central metabolic pathway in escherichia coli to reduce acetate accumulation by heterologous expression of the bacillus subtilis acetolactate synthase gene. Biotechnol. Bioeng. 44, 944–951. https://doi.org/10.1002/bit.260440810.

Bi, X., Liu, L.F., 1996. DNA rearrangement mediated by inverted repeats. Proc. Natl. Acad. Sci. U. S. A 93, 819–823. https://doi.org/10.1073/pnas.93.2.819.

Borkowski, O., Ceroni, F., Stan, G.B., Ellis, T., 2016. Overloaded and stressed: whole-cell considerations for bacterial synthetic biology. Curr. Opin. Microbiol. 33, 123–130. https://doi.org/10.1016/j.mib.2016.07.009.

Bull, J.J., Barrick, J.E., 2017. Arresting evolution. Trends Genet. 33, 910–920. https://doi.org/10.1016/j.tig.2017.09.008.

Ceroni, F., Furini, S., Gorochowski, T.E., Boo, A., Borkowski, O., Ladak, Y.N., Awan, A.R., Gilbert, C., Stan, G.-B., Ellis, T., 2018. Burden-driven feedback control of gene expression. Nat. Methods 1–17. https://doi.org/10.1101/177030.

Chang, A.C.Y., Cohen, S.N., 1978. Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. J. Bacteriol. 134, 1141–1156. https://doi.org/10.1128/jb.134.3.1141-1156.1978.

Cox, R., Mirkin, S.M., 1997. Characteristic enrichment of DNA repeats in different genomes. Proc. Natl. Acad. Sci. U. S. A 94, 5237–5242. https://doi.org/10.1073/pnas.94.10.5237.

D'Ambrosio, V., Dore, E., Di Blasi, R., van den Broek, M., Sudarsan, S., Horst, J. ter, Ambri, F., Sommer, M.O.A., Rugbjerg, P., Keasling, J.D., Mans, R., Jensen, M.K., 2020. Regulatory control circuits for stabilizing long-term anabolic product formation in yeast. Metab. Eng. 61, 369–380. https://doi.org/10.1016/j.ymben.2020.07.006.

Darmon, E., Leach, D.R.F., 2014. Bacterial genome instability. Microbiol. Mol. Biol. Rev. 78, 1–39. https://doi.org/10.1128/MMBR.00035-13.

Deatherage, D.E., Traverse, C.C., Wolf, L.N., Barrick, J.E., 2015. Detecting rare structural variation in evolving microbial populations from new sequence junctions using breseq. Front. Genet. 5, 1–16. https://doi.org/10.3389/fgene.2014.00468.

Durfee, T., Nelson, R., Baldwin, S., Plunkett, G., Burland, V., Mau, B., Petrosino, J.F., Qin, X., Muzny, D.M., Ayele, M., Gibbs, R.a., Csorgo, B., Posfai, G., Weinstock, G.M., Blattner, F.R., 2008. The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. J. Bacteriol. 190, 2597–2606. https://doi.org/10.1128/JB.01695-07.

Fan, C., Wu, Y.H., Decker, C.M., Rohani, R., Gesell Salazar, M., Ye, H., Cui, Z., Schmidt, F., Huang, W.E., 2019. Defensive function of transposable elements in bacteria. ACS Synth. Biol. 8, 2141–2151. https://doi.org/10.1021/acssynbio.9b00218.

Geng, P., Leonard, S.P., Mishler, D.M., Barrick, J.E., 2019. Synthetic genome defenses against selfish DNA elements stabilize engineered bacteria against evolutionary failure. ACS Synth. Biol. 8, 521–531. https://doi.org/10.1021/acssynbio.8b00426.

Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S., Ohtsubo, E., Baba, T., Wanner, B.L., Mori, H., Horiuchi, T., 2006. Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. Mol. Syst. Biol. 2 https://doi.org/10.1038/msb4100049.

Ikeda, M., 2003. Amino acid production processes. Adv. Biochem. Eng. Biotechnol. 79, 1–35. https://doi.org/10.1007/3-540-45989-8_1.

Jeong, H., Barbe, V., Lee, C.H., Vallenet, D., Yu, D.S., Choi, S.H., Couloux, A., Lee, S.W., Yoon, S.H., Cattolico, L., Hur, C.G., Park, H.S., Ségurens, B., Kim, S.C., Oh, T.K., Lenski, R.E., Studier, F.W., Daegelen, P., Kim, J.F., 2009. Genome sequences of Escherichia coli B strains REL606 and BL21(DE3). J. Mol. Biol. 394, 644–652. https://doi.org/10.1016/j.jmb.2009.09.052.

Jiang, X., Hall, A.B., Arthur, T.D., Plichta, D.R., Covington, C.T., Poyet, M., Crothers, J., Moses, P.L., Tolonen, A.C., Vlamakis, H., Alm, E.J., Xavier, R.J., 2019. Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. Science 363, 181–187. https://doi.org/10.1126/science.aau5238.

Jinks-Robertson, S., Bhagwat, A.S., 2014. Transcription-associated mutagenesis. Annu. Rev. Genet. 48, 341–359. https://doi.org/10.1146/annurev-genet-120213-092015.

Kim, J.F., Jeong, H., Park, J., Vickers, C.E., Lee, S.Y., Nielsen, L.K., 2011. The genome sequence of *E. coli* W (ATCC 9637): Comparative genome analysis and an improved genome-scale reconstruction of *E. coli*. BMC Genom. https://doi.org/10.1186/1471-2164-12-9.

Kim, B., Park, H., Na, D., Lee, S.Y., 2014. Metabolic engineering of Escherichia coli for the production of phenol from glucose. Biotechnol. J. 9, 621–629. https://doi.org/10.1002/biot.201300263.

Kizer, L., Pitera, D.J., Pfleger, B.F., Keasling, J.D., 2008. Application of functional genomics to pathway optimization for increased isoprenoid production. Appl. Environ. Microbiol. 74, 3229–3241. https://doi.org/10.1128/AEM.02750-07.

Kwon, S.K., Kim, S.K., Lee, D.H., Kim, J.F., 2015. Comparative genomics and experimental evolution of Escherichia coli BL21(DE3) strains reveal the landscape of toxicity escape from membrane protein overproduction. Sci. Rep. 5, 16076. https://doi.org/10.1038/srep16076.

Lavi, B., Karin, E.L., Pupko, T., Hazkani-Covo, E., 2018. The prevalence and evolutionary conservation of inverted repeats in proteobacteria. Genome Biol. Evol. 10, 918–927. https://doi.org/10.1093/gbe/evy044.

Lv, Y., Gu, Y., Xu, J., Zhou, J., Xu, P., 2020. Coupling metabolic addiction with negative autoregulation to improve strain stability and pathway yield. Metab. Eng. 61, 79–88. https://doi.org/10.1016/j.ymben.2020.05.005.

Martin, V.J.J., Pitera, D.J., Withers, S.T., Newman, J.D., Keasling, J.D., 2003. Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. Nat. Biotechnol. 21, 796–802. https://doi.org/10.1038/nbt833.

Monk, J.M., Koza, A., Campodonico, M.A., Machado, D., Seoane, J.M., Palsson, B.O., Herrgård, M.J., Feist, A.M., 2016. Multi-omics quantification of species variation of Escherichia coli links molecular features with strain phenotypes. Cell Syst. 3, 238–251. https://doi.org/10.1016/j.cels.2016.08.013 e12.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N., Kanaya, S., 2011. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 39 https://doi.org/10.1093/nar/gkr344.

Nielsen, J., Keasling, J., 2016. Engineering cellular metabolism. Cell 164, 1185–1197.

Nyerges, Á., Bálint, B., Cseklye, J., Nagy, I., Pál, C., Fehér, T., 2019. CRISPR-interference based modulation of mobile genetic elements in bacteria. Synth. Biol. 1–39 https://doi.org/10.1093/synbio/ysz008.

Park, J.H., Jang, Y.S., Lee, J.W., Lee, S.Y., 2011. Escherichia coli W as a new platform strain for the enhanced production of L-Valine by systems metabolic engineering. Biotechnol. Bioeng. 108, 1140–1147. https://doi.org/10.1002/bit.23044.

Payen, C., Di Rienzi, S.C., Ong, G.T., Pogachar, J.L., Sanchez, J.C., Sunshine, A.B., Raghuraman, M.K., Brewer, B.J., Dunham, M.J., 2014. The dynamics of diverse segmental amplifications in populations of Saccharomyces cerevisiae adapting to strong selection. G3 Genes, Genom. Genet. 4, 399–409. https://doi.org/10.1534/g3.113.009365.

Quick, J., Grubaugh, N.D., Pullan, S.T., Claro, I.M., Smith, A.D., Gangavarapu, K., Oliveira, G., Robles-Sikisaka, R., Rogers, T.F., Beutler, N.A., Burton, D.R., Lewis-Ximenez, L.L., De Jesus, J.G., Giovanetti, M., Hill, S.C., Black, A., Bedford, T., Carroll, M.W., Nunes, M., Alcantara, L.C., Sabino, E.C., Baylis, S.A., Faria, N.R., Loose, M., Simpson, J.T., Pybus, O.G., Andersen, K.G., Loman, N.J., 2017. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat. Protoc. 12, 1261–1266. https://doi.org/10.1038/nprot.2017.066.

Rugbjerg, P., Myling-Petersen, N., Porse, A., Sarup-Lytzen, K., Sommer, M.O.A., 2018a. Diverse genetic error modes constrain large-scale bio-based production. Nat. Commun. 9, 787. https://doi.org/10.1038/s41467-018-03232-w.

Rugbjerg, P., Olsson, L., 2020. The future of self-selecting and stable fermentations. J. Ind. Microbiol. Biotechnol. 0000-0003-2561-5063.

Rugbjerg, P., Sarup-Lytzen, K., Nagy, M., Sommer, M.O.A., 2018b. Synthetic addiction extends the productive life time of engineered Escherichia coli populations. Proc. Natl. Acad. Sci. Unit. States Am. 115, 2347–2352. https://doi.org/10.1073/pnas.1718622115.

Rugbjerg, P., Sommer, M.O.A., 2019. Overcoming genetic heterogeneity in industrial fermentations. Nat. Biotechnol. 37, 869–876. https://doi.org/10.1038/s41587-019-0171-6.

Sandberg, T.E., Salazar, M.J., Weng, L.L., Palsson, B.O., Feist, A.M., 2019. The emergence of adaptive laboratory evolution as an efficient tool for biological discovery and industrial biotechnology. Metab. Eng. 56, 1–16. https://doi.org/10.1016/j.ymben.2019.08.004.

Sedlazeck, F.J., Lee, H., Darby, C.A., Schatz, M.C., 2018a. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat. Rev. Genet. 19 (6), 329–346. https://doi.org/10.1038/s41576-018-0003-4.

Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., Schatz, M.C., 2018b. Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods 15, 461–468. https://doi.org/10.1038/s41592-018-0001-7.

Sousa, A., Bourgard, C., Wahl, L.M., Gordo, I., 2013. Rates of transposition in Escherichia coli. Biol. Lett. 9, 2–5. https://doi.org/10.1098/rsbl.2013.0838.

Takors, R., 2012. Scale-up of microbial processes: impacts , tools and open questions. J. Biotechnol. 160, 3–9. https://doi.org/10.1016/j.jbiotec.2011.12.010.

Tyo, K.E.J., Ajikumar, P.K., Stephanopoulos, G., 2009. Stabilized gene duplication enables long-term selection-free heterologous pathway expression. Nat. Biotechnol. 27, 760–765. https://doi.org/10.1038/nbt.1555.

Umenhoffer, K., Draskovits, G., Nyerges, A., Karcagi, I., Bogos, B., Tímár, E., Csörgö, B., Herczeg, R., Nagy, I., Fehér, T., Pál, C., Pósfai, G., 2017. Genome-wide abolishment of mobile genetic elements using genome shuffling and CRISPR/Cas-Assisted MAGE allows the efficient stabilization of a bacterial chassis. ACS Synth. Biol. 6, 1471–1483. https://doi.org/10.1021/acssynbio.6b00378.

Wehrs, M., Tanjore, D., Eng, T., Lievense, J., Pray, T.R., Mukhopadhyay, A., 2019. Engineering robust production microbes for large-scale cultivation. Trends Microbiol. 27, 524–537. https://doi.org/10.1016/j.tim.2019.01.006.

Xiong, M., Schneiderman, D.K., Bates, F.S., Hillmyer, M. a, Zhang, K., 2014. Scalable production of mechanically tunable block polymers from sugar. Proc. Natl. Acad. Sci. U. S. A 111, 8357–8362. https://doi.org/10.1073/pnas.1404596111.

Xu, P., 2018. Production of chemicals using dynamic control of metabolic fluxes. Curr. Opin. Biotechnol. 53, 12–19. https://doi.org/10.1016/j.copbio.2017.10.009.

Xu, Y., Chu, H., Gao, C., Tao, F., Zhou, Z., Li, K., Li, L., Ma, C., Xu, P., 2014. Systematic metabolic engineering of Escherichia coli for high-yield production of fuel biochemical 2,3-butanediol. Metab. Eng. 23, 22–33. https://doi.org/10.1016/j.ymben.2014.02.004.

Zelder, O., Hauer, B., 2000. Environmentally directed mutations and their impact on industrial biotransformation and fermentation processes. Curr. Opin. Microbiol. 3, 248–251. https://doi.org/10.1016/S1369-5274(00)00084-9.