Tracking the roots of cellulase hyperproduction by the fungus *Trichoderma reesei* using massively parallel DNA sequencing

Stéphane Le Crom[a,b,c,1], Wendy Schackwitz[d,1], Len Pennacchio[d], Jon K. Magnuson[e], David E. Culley[e], James R. Collett[e], Joel Martin[d], Irina S. Druzhinina[f], Hugues Mathis[g], Frédéric Monot[g], Bernhard Seiboth[f], Barbara Cherry[h], Michael Rey[h], Randy Berka[h], Christian P. Kubicek[f], Scott E. Baker[d,e,2], and Antoine Margeot[g,2]

[a]Institut National de la Santé et de la Recherche Médicale, U784, 46 rue d'Ulm, 75230 Paris Cedex 05, France; [b]Institut Fédératif de Recherche 36, Plate-forme Transcriptome, 46 rue d'Ulm, 75230 Paris Cedex 05, France; [c]École Normale Supérieure, 46 rue d'Ulm, 75230 Paris Cedex 05, France; [d]Department of Energy Joint Genome Institute, 2800 Mitchell Avenue, Walnut Creek, CA 94598; [e]Pacific Northwest National Laboratory, P.O. Box 999, Richland, WA 99352; [f]Institute of Chemical Engineering, Technische Universitat Wien, Getreidemarkt 9/166, A-1060 Vienna, Austria; [g]IFP, Département Biotechnologie, Avenue de Bois-Préau, 92852 Rueil-Malmaison Cedex, France; and [h]Novozymes, Inc., 1445 Drew Avenue, Davis, CA 95618

Edited by Joan Wennstrom Bennett, Rutgers University, New Brunswick, NJ, and approved July 27, 2009 (received for review May 28, 2009)

Trichoderma reesei **(teleomorph** ***Hypocrea jecorina*****) is the main industrial source of cellulases and hemicellulases harnessed for the hydrolysis of biomass to simple sugars, which can then be converted to biofuels such as ethanol and other chemicals. The highly productive strains in use today were generated by classical mutagenesis. To learn how cellulase production was improved by these techniques, we performed massively parallel sequencing to identify mutations in the genomes of two hyperproducing strains (NG14, and its direct improved descendant, RUT C30). We detected a surprisingly high number of mutagenic events: 223 single nucleotides variants, 15 small deletions or insertions, and 18 larger deletions, leading to the loss of more than 100 kb of genomic DNA. From these events, we report previously undocumented non-synonymous mutations in 43 genes that are mainly involved in nuclear transport, mRNA stability, transcription, secretion/vacuolar targeting, and metabolism. This homogeneity of functional categories suggests that multiple changes are necessary to improve cellulase production and not simply a few clear-cut mutagenic events. Phenotype microarrays show that some of these mutations result in strong changes in the carbon assimilation pattern of the two mutants with respect to the wild-type strain QM6a. Our analysis provides genome-wide insights into the changes induced by classical mutagenesis in a filamentous fungus and suggests areas for the generation of enhanced** ***T. reesei*** **strains for industrial applications such as biofuel production.**

biofuels | biotechnology

Even 25 years after the invention of recombinant technologies, many biotechnological processes still make use of microbial strains that have been improved by classical mutagenesis. Information about the loci that became altered in the process of mutation and selection for improved product titers is scarce, if available at all, due to the lack of availability of tractable methods for their discovery. However, with the advent of new high-throughput, massively parallel sequencing technologies, an accurate characterization of a mutant genome relative to a previously sequenced parental reference strain has become a realistic approach. In fact, a recent study with the 15.4 Mb genome of the yeast *Pichia stipitis* that compared three sequencing technologies (Life Sciences, Roche; Illumina sequencing; and Applied Biosystems SOLiD) found that they were all able to consistently identify a common set of single nucleotide variants assuming a minimum of 10–15-fold nominal sequence coverage (1).

Trichoderma reesei (teleomorph *Hypocrea jecorina*) is the workhorse organism for a number of industrial enzyme companies for the production of cellulases (2–4). Using random mutagenesis, academic and industrial research programs have over several decades produced strains of *T. reesei* whose production of cellulases is several times higher than that of the "original" *T. reesei* strain QM6a that was isolated from US Army tent canvas in 1944 in the Solomon Islands (5). Despite these years of research and development in the area, the costs associated with production of enzymes that degrade biomass are still considered a significant barrier to economic lignocellulosic fuel ethanol. The genetics underlying the respective phenotypes of these mutants is essentially unknown. Understanding the molecular mechanisms that underlie the improvements made by random mutagenesis of *T. reesei* QM6a could open avenues for construction of better and more efficient cellulase producing strains by targeted molecular genetic manipulation.

One of the best producer strains in the public domain is *T. reesei* RUT C30 (6). This mutagenized strain was obtained through three mutagenesis steps. First, UV mutagenesis, followed by selection for the ability to hydrolyze cellulose in catabolite repressing conditions, led to strain M7 (this strain is no longer available). Second, strain NG14 was derived from M7 through chemical (N-nitrosoguanidine, NTG) mutagenesis using a similar but more stringent screen. NG14 exhibited several-fold increases in extracellular protein and cellulase activity compared with parental strains and other cellulase mutants that were available (6). Strain RUT C30 was produced from NG14 using UV mutagenesis and was screened with a similar cellulose hydrolysis assay and for resistance to 2-deoxyglucose to eliminate catabolite repression (7). Accumulation of 2-deoxyglucose-6-phosphate rapidly leads to growth inhibition (8). The resulting strain produces twice as much extracellular protein relative to its parental strain NG14, reaching more than 30 g/L production in industrial fermentations and also exhibited catabolite derepression (6). The genealogy of the strains is presented in Fig. 1.

In the years following its generation, RUT C30 has become a reference strain among *T. reesei* high cellulase producers, and it has been used in numerous studies (9, 10). Electrophoretic karyotyping of RUT C30 (11, 12) revealed chromosomal rear-

Author contributions: C.P.K., S.E.B., and A.M. designed research; S.L.C., W.S., L.P., J.K.M., D.E.C., J.R.C., J.M., I.S.D., H.M., F.M., B.C., and M.R. performed research; S.L.C., W.S., I.S.D., B.S., B.C., M.R., R.B., C.P.K., S.B., and A.M. analyzed data; and S.L.C., W.S., R.B., C.P.K., S.E.B., and A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The complete list of accession numbers reported in this work is available in the supplemental information Table S5 and at www.ncbi.nlm.gov/snp_viewBatch.cgi?sbid=1048820.

[1]S.L.C. and W.S. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: antoine.margeot@ifp.fr or scott.baker@pni.gov.

This article contains supporting information online at www.pnas.org/cgi/content/full/0905848106/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0905848106

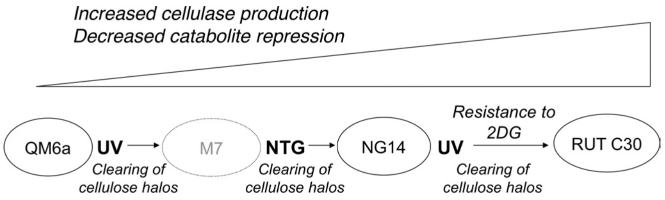
Increased cellulase production
Decreased catabolite repression
QM6a
UV
Clearing of cellulose halos
M7
NTG
Clearing of cellulose halos
NG14
UV
Resistance to 2DG
Clearing of cellulose halos
RUT C30

Fig. 1. Genealogy of strains used in this study. Mutagens used appear in bold next to strain names. Screening procedures are indicated in italics. 2DG stands for 2-deoxy-glucose. The gray color used for the M7 strain indicates the strain is no longer available, and could not be included in this study.

rangements but precise genetic changes that occurred in the strain are poorly characterized. To date, three mutations in *T. reesei* RUT C30 have been uncovered: a truncation of the *cre1* gene (tre120117), a key carbon catabolite repression mediator (13); a frameshift mutation in the glucosidase II alpha subunit gene *gls2α* (tre121351) involved in protein glycosylation and whose replacement with the wild-type allele decreases protein secretion (14); and an 85-kb deletion that eliminated 29 genes, including transporters, transcription factors, and primary metabolic enzymes (15). While the QM6a to RUT C30 lineage is not the only one available, it remains the most popular, the one with the highest number of genomic variations reported, and the only one with direct ancestry on the hyperproducer strain available. It is an obvious choice to build a genome-wide study on solid foundations.

To better understand the biology underlying improvements to *T. reesei* high cellulase-secreting strains, we have used massively parallel sequencing to characterize the genomes of two strains that were derived from the wild-type *T. reesei* QM6a: NG14 and RUT C30. Changes in genome composition were compared with the recently published *T. reesei* QM6a genome sequence (16). The strains were also assayed with Biolog carbon source phenotype arrays to assess how carbon assimilation profiles have been altered as a consequence of strain selection and gene modifications.

Results

Sequencing of NG14 and RUT C30 *T. reesei* Strains. To uncover genetic changes that occurred between *T. reesei* strains QM6a, NG14, and RUT C30, we implemented a massively parallel sequencing approach with the Illumina Solexa technology. To avoid bias due to sequencing methods or genetic drift of the strains, two independent isolates of the RUT C30 strain were sequenced and analyzed. Only one NG14 isolate was sequenced. For the first RUT C30 isolate we mapped 23,965,578 (94.1% of total) single end reads with an average depth of 25.7 and 35,783,984 (94.9% of total) paired end reads (52% with 3.3-kb inserts, 26% with 300-bp inserts, and 22% with misoriented/chimeric inserts) with an average depth of 33.4 for the second isolate. For NG14, we mapped 16,165,618 single end reads (71.4% of total) with an average depth of 16.4. We subsequently analyzed the data for four kinds of mutational events: single nucleotide variants (SNVs), small deletions and insertions (indels), large deletions, and duplication events.

Single Nucleotide Variants. Our simulated SNV analysis (see Materials and Methods section) identified approximately 90% of all SNVs for NG14 and approximately 97–98% for the two RUT C30 isolates, suggesting a nearly exhaustive coverage at least for RUT C30. In addition to the quality filter that was applied to SNVs described in the Materials and Methods section, we performed an additional filtering step using the genomic context information in a 60-base window around each SNV. We kept SNVs with a complexity score of 1, a GC percentage in the 60-base window between 31% and 74%, and a uniqueness score greater than 15.8. These thresholds were fixed using the mean value of each parameter distribution and two times the standard deviation. Using these filters we identified 103 SNVs for NG14 and 220 for RUT C30. Based on the known history of NG14 and the two RUT RUT C30 isolates, there were 30 SNVs that were inconsistent, 21 present in one RUT C30 isolate and not in the other, and nine present in NG14 only. These SNVs were referred to as "orphans." To ensure that these orphans were indeed absent in the other strain/isolate and not false negatives due to low local coverage or low quality sequences, we manually inspected the raw sequence data for NG14 and the two RUT C30 isolates. This allowed us to validate that 20 of the orphans had been false negatives. Four more SNVs were confirmed as NG14-specific and were attributed to genetic drift independent from strain selections. Five SNVs could not be confirmed in both RUT C30 strains and were thus eliminated. Additionally, 33 SNVs identified in RUT C30 that had been ruled out because of low coverage in NG14 were manually confirmed by looking at raw data. In all we identified 136 SNVs in *T. reesei* NG14 relative to QM6a, and an additional 99 specifically in RUT C30. Sanger sequencing of 120 randomly chosen SNV positions in the QM6a strain led to the elimination of 12 SNVs that were in fact errors in the original QM6a sequence. The final list of SNVs identified and validated in NG14 and RUT C30 is available in Table S1.

Small Deletion and Insertion Events. For the RUT C30 isolate that was paired-end sequenced we used Maq's indel detection function to identify small insertions and deletions (indels) up to 6bp long. Since Maq's indel detection only functions with small inserts 26% of the paired end data (9× average depth) was used to detect small indels. Using the simulated indels method, we estimated that we have identified 68% of all indels <7bp (see *Materials and Methods* section). Indel positions determined were then manually inspected and validated in NG14 and other RUT C30 isolate. As for the SNVs, only consistent indels taking into account strain history were conserved. In all, 15 indels were validated: 11 in both strains, and four in RUT C30 alone. The complete list of identified indels is available in Table S2.

Large Deletions and Duplications. To detect large deletions and duplications we applied a sliding window strategy to calculate the moving average of the coverage along the sequenced genomes. We choose a 200-bp window size as this is the breakpoint in both strains in coverage variance stabilization. For deletion detection we fixed the coverage threshold to one. As for SNVs, we calculated the average genomic context information along each deletion and we filtered out deletions where more than 75% of the region is unknown in QM6a and with a complexity score below 0.75. With these filters, we detected 11 deletions in NG14 and seven more in RUT C30. The smallest deletion identified using this strategy encompassed about 90 bp (see Table S3).

We used the same the same 200-bp window to detect large duplications. We set the detection threshold to mean coverage plus one standard deviation, which gave 32 for NG14 and 42 for RUT C30. Using this method in combination with the same genomic context filters as above, we could not identify putative amplifications in any of the three datasets.

Mutation Patterns in the Two Strains. The majority of mutational events were single nucleotide changes, of which 126 were found in *T. reesei* NG14, and an additional 97 specifically occurred only in RUT C30 leading to a total of 223 SNVs in this strain (See Table 1). Interestingly, the identified single nucleotide changes caused by the mutations were significantly different between *T. reesei* NG14 and RUT C30 (Fig. 2). Thirty percent of the nucleotide changes in strain NG14 were A-T→G-C or, while these accounted for only 9% of RUT C30 specific mutations. In RUT C30, 75% of the changes were G-C→A-T. These values

Table 1. SNV distribution and chromosomal features affected in sequenced strains

	From QM6a to NG14	From NG14 to RUT C30	Total in RUTC30
SNV distribution in sequenced strains			
Total SNVs	126	97	223
In promoters	46	21	66
In terminators	10	2	13
Elsewhere	48	29	77
Total intergenic hits*	103	52	155
In introns	10	6	16
In exons	24	43	67
Synonymous mutations	5	18	23
Nonsynonymous mutations	19	25	44
Distinct genetic elements affected			
Promoters	42	17	59
Terminators	9	2	11
Introns	8	6	14
CDS	21	41	62
Mutated proteins	17	25	42

*The differences in summing take into account mutations affecting at the same time two features, such as a promoter and a terminator. Differences with figures from SNV distribution in sequenced strains take into account two mutations affecting a single genetic feature. Differences in nonsilent mutations and mutated proteins are due to several single-nucleotide exchanges occurring in the same gene.

coincide well with the use of the two different mutagens (NTG and UV light) for NG14, and then UV light alone for RUT C30. Interestingly, NTG mutagenesis is reported to lead mainly to A-T→G-C mutations (17), which is consistent with patterns observed here for NG14.

Nineteen of the 24 point mutations (79%) that mapped to exons in strain NG14 and 25 of 43 (58%) that mapped to exons in RUT C30 were non-synonymous, and thus, potentially affect the properties of the corresponding proteins. Remarkably, the number of mutations resulting in protein modification that occurred during the RUT C30 mutagenesis step (25%) was significantly higher than in the NG14 step (15%), despite a higher total number of SNVs from the QM6a to NG14 mutagenesis (Table 1). This may be due to the different physiological conditions used for selection of the strains (i.e., resistance to 2-deoxyglucose for RUT C30, see Discussion). A number of mutations in promoters (45 in NG14 and 21 in RUT C30) and terminators (11 in NG14 and 2 in RUT C30) were also identified (Table 1 and Table S1).

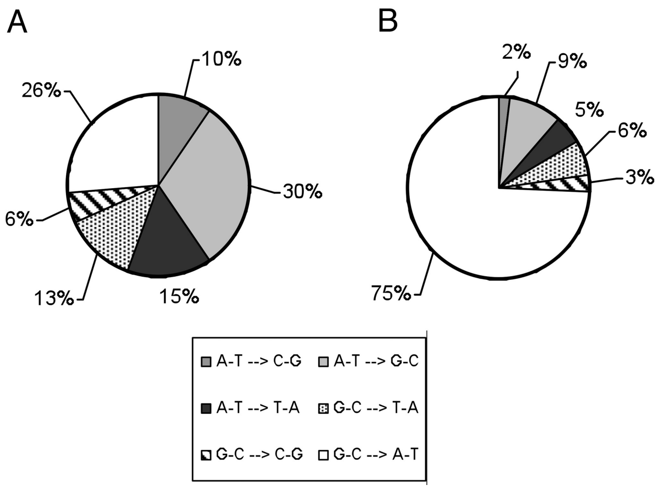

Fig. 2. Nucleotide exchanges found in the SNVs in *T. reesei* NG14 (*A*) and *T. reesei* RUT C30 (*B*). Percentages were calculated based on the 126 exchanges in NG14 and 97 further exchanges in RUT C30. The more diversified mutation pattern in NG14 reflects the diversity of mutagen used (UV and NTG), to be compared to UV alone for RUT C30.

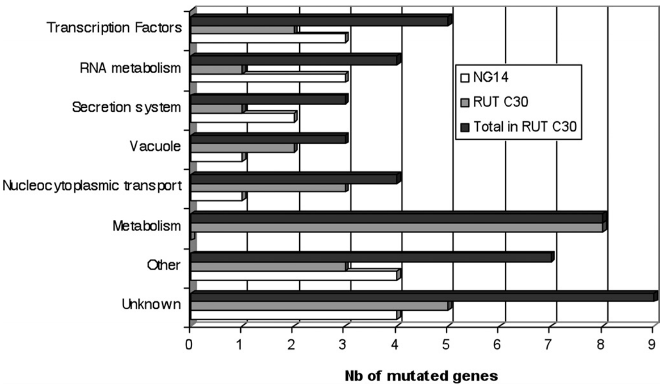

Fig. 3. Gene categories affected by mutagenic events in both strains. White and gray bars respectively indicate the number of mutations for a specific gene category observed for NG14 mutagenesis and for RUT C30 mutagenesis. The black bars indicate total mutations found in RUT C30 including the ones from the NG14 mutagenesis round. The full list of genes for each category can be found in Table S4.

Of the 15 indels detected, two resulted in frameshift mutations, including one already reported: a single T deletion resulting in a frameshift and truncation in the *gls2α* gene encoding a processing ß-glucosidase subunit (14). This frameshift was found in both NG14 and RUT C30. Another deletion (−TCCC) at the end of gene model tre3400 also ostensibly gives rise to an abbreviated gene product. Two previously reported large deletions were confirmed and appear to be the major deletion events in these strains. The first is an 85-kb deletion on Scaffold 15, which is already present in *T. reesei* NG14, resulting in the loss of 29 genes (15). The second is the truncation of the carbon catabolite repressor gene *cre1* in RUT C30, which renders this strain partially carbon catabolite derepressed (13, 15). Seventeen other deletions, with sizes ranging from 0.91 to 5.2 kb were also identified. Only two of them hit identified genetic elements: one hits the 5′ part of tre120806 coding sequence, and the other one hits a promoter (see Table S3).

Genes Affected. We then considered genes affected by the identified mutational events: 18 in NG14 and 25 more in RUT C30. The loci and (putative) functions are listed in Table S4. Half of the 18 mutated genes in *T reesei* NG14 is involved in either RNA metabolism (3 genes), in protein secretion and vacuolar targeting (3 genes), or encoding transcription factors (3 genes) (Fig. 3). Interestingly, additional genes belonging to the same categories were further mutated in *T. reesei* RUT C30 (1, 3, and 2 genes, respectively), further highlighting components of apparently major importance to cellulase hyperproduction. In addition, RUT C30 has accumulated a considerable number of mutations in genes involved in sugar transport and general metabolism (8 genes), which may be related to high selective-pressure selection for growth on glycerol in the presence of 2-deoxyglucose (see *Discussion*).

An interesting finding in this context is the accumulation of several mutations in a gene encoding a putative kinase tre120806 (two SNVs and a C-terminal deletion) and in a gene encoding a kinesin related protein tre112231 (two SNVs), as well as several genes associated with the same processes: a mutation in glycerol-

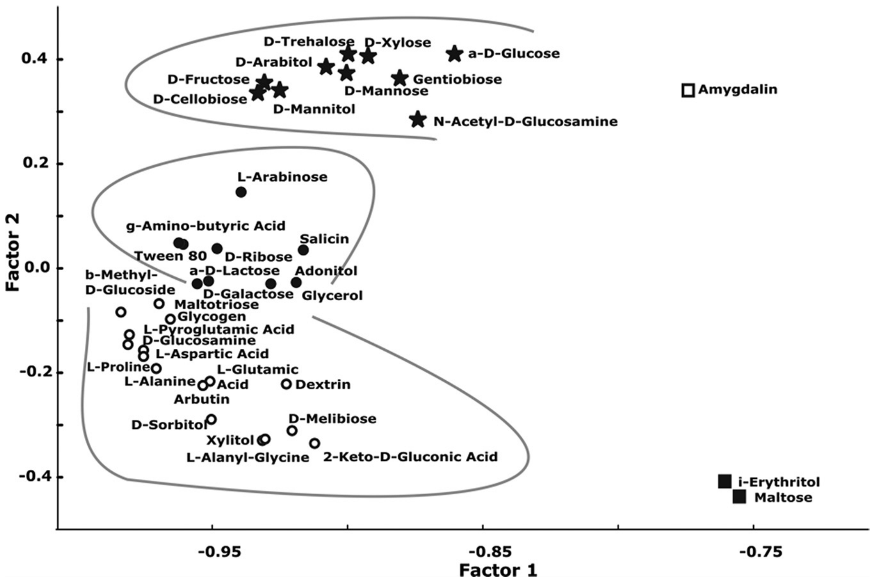

Fig. 4. Results of factor analysis applied to changes in biomass concentration during the linear growth phase. Only carbon sources supporting good growth of QM6a [Clusters I and II, (37)] are included. Stars correspond to carbon sources which supported superior growth of QM6a and RUT C30 compared to QM6a; open cycles correspond to carbon sources on which both mutants had reduced growth compared to the wild-type strain; filled cycles correspond to those carbon sources on which RUT C30 had the reduced growth compared to NG14 while both mutants grew slower compared to QM6a. Open squares indicate the only case of superior growth of NG14; filled squares correspond to erythritol and maltose. On the first carbon source, both mutant strains had no growth compared to the wild-type. On the later one, only NG14 could not grow while RUT C30 showed the recovered phenotype of the wild-type strain.

3-phosphate phosphatase (tre58790) together with the loss of glycerol dehydrogenase by the 85-kb gap, and the mutation in the maltose permease (tre298667) together with the loss of a maltose permease by the 85-kb gap (15). Several nucleotide changes were also noted in the promoter regions of genes (Table S1).

Carbon Source Assimilation Profiles. Following mutagenesis *T. reesei* strains NG14 and RUT C30 were identified on the basis of their increased cellulase activity, and (in the latter case) resistance to 2-deoxyglucose in the presence of glycerol. Since the data reported above suggest a considerable number of mutations accumulated in these strains, we wondered whether they may have additional phenotypic differences besides increased cellulase production. To this end, we used Biolog Phenotype microarrays containing 95 different carbon sources. Growth of *T reesei* QM6a and the two mutants was monitored in triplicate over a period of 96 h. Specific growth rates were calculated and subjected to factorial analyses (Fig. 4). Carbon sources, whose utilization correlated with increased cellulase production included glucose, fructose mannose, N-acetylglucosamine and trehalose, D-xylose, D-arabinitol, mannitol, and the ß-linked disaccharides gentiobiose and cellobiose. While the latter two might reflect the increase in ß-glucosidase activity for which these two are substrates, it is notable that most of the other compounds are catabolite repressing carbon sources. Whether this is solely a consequence of the *cre1* mutation or is due to a combination of the other genes affected remains to be investigated.

Utilization of another set of carbon sources correlated inversely with cellulase production. These included the utilization of α–linked oligosaccharides and glycans observed previously (15), and which are likely due to the loss of the maltose permease gene in the 85-kb deletion on Scaffold 15 and the mutation in the other maltose uptake gene (tre298667) reported above. Additionally, we observed that increased cellulase productivity was correlated with reduced growth on amino acids. Reduced growth on amino acids may at a first glance seem to be contradictory to higher cellulase (and thus protein) production, but this phenotype may be related to an ability of the improved strains to use a higher portion of their amino acid pool for synthesis of secreted proteins versus growth.

Interestingly—and in contrast to its importance as an alternative inducer of cellulase formation—the utilization of lactose and its constituent D-galactose decreased with increasing cellulase production. This suggests that the rate limiting step(s) in cellulase induction by cellulose (which was exclusively used for screening these mutants) might be different from those affected by lactose, or that function of the latter as inducer is indirectly correlated with its utilization as a carbon source.

Discussion

Although the overwhelming majority of filamentous fungi used in industry have been generated by classical mutagenesis, there are only a few cases where the genes affected by these mutations have been identified. This study is a comprehensive whole genome analysis for two important strains of *T. reesei*, produced in direct series from the wild-type strain. In addition to confirming previously reported mutations (13–15), we also identified many genomic alterations including 223 SNVs, 15 small deletions or insertions, and 18 larger deletions (>90 bp). Using 34.1 Mbp as the genome size of *T. reesei* (16), the observed number of SNPs corresponds to a mutation rate of 6.56×10^{-6} per whole genome, which is approximately 10^3 fold higher than the spontaneous mutation rate. The round of mutagenesis leading from *T. reesei* NG14 to RUT C30 introduced proportionally more SNVs in exons (44% versus 19% in NG14), resulting in a higher number of genes affected by amino acid changes (25 versus 17 in NG14) than in the first two rounds. When the fraction of mutations in ORFs was compared to the average gene density [40.40%; (16)], NG14 clearly fell short (19%) whereas RUT C30 was slightly above this value (44%). Noteworthy is also the difference between the percentage of non-synonymous mutations retained in the two strains (79%; 19 of 24 in NG14 and 58%; 25 of 43 in RUT C30). This may be due to the different physiological conditions used for selection of the strains. Selection could have been more severe for RUT C30 resulting in positive selection and thus preferential retention of non-synonymous SNVs. Alternatively, higher negative-selective pressure on the NG14 strain after mutagenesis could have resulted in the sweeping out of deleterious SNVs in NG14. Another possible interpretation is that this higher mutation rate was due to relief of selective pressure on these genes after mutation of key genes for cellulase production in the mutagenesis steps that led to NG14.

A key feature in the selection of *T. reesei* RUT C30 was growth on glycerol as a carbon source in the presence of 2-deoxyglucose (2-DG). The use of this analog probably facilitated the isolation of a mutant that contained a deletion in the carbon catabolite repressor CRE1-encoding *cre1* gene (13). Ralser et al. (18) have recently shown that resistance to 2-deoxyglucose in *S. cerevisiae* involves several other metabolic traits such as mitochondrial homeostasis, mRNA decay, transcriptional regulation and the cell cycle. It is therefore tempting to speculate that some of the genes involved in these gene categories which are affected in RUT C30 may be also related to 2-deoxyglucose resistance. In fact, the mutation in glycerol-3-phosphate phosphatase (tre58790) may have been beneficial to combat the effect of 2-DG in the glycerol dehydrogenase-negative background present in NG14 (15).

In addition to the 29 genes removed by the large 85-kb deletion which occurred during the generation of *T. reesei* NG14, the truncation of *cre1* gene and the frameshift in glucosidase II, an additional 18 CDSs in NG14, and 25 more in RUT C30 bear non-synonymous SNVs, or are affected by deletions. While the mere presence of SNVs does not allow us to conclude whether they have a biologically significant effect, it is intriguing that there is a clear bias in the genes involved. Nearly 45% of the genes affected encoded transcription factors, components of nuclear import, mRNA metabolism, protein secretion, and vacuolar sorting. Transcription factors were most abundant among them. When *cre1* (tre120117) and the four proteins which were lost by the 85-kb deletion (15) were included (tre44995, tre36620, tre79725, and tre65070), a total of 10 putative transcription factors were identified. Nine belong to the fungal-specific Zn (2)-Cys (6)-class of zinc finger proteins (19) and showed putative orthologues in other fungi although none has been characterized. Given the fact that fungal genomes typically contain 200 to 300 transcription factors of this class, their mutation frequency calculates as 3–4.5 $\times 10^{-2}$, which is significantly higher than the mutation rate of 6.56 $\times 10^{-6}$ per whole genome calculated above. We therefore consider it unlikely that they are unrelated to the selection process and thus cellulase production. The other, a CBF/NF-Y family transcription factor termed negative cofactor B2 (tre109619), is the beta subunit of a negative regulator of the RNA polymerase II holoenzyme, and it shares sequence identity with the Dr1 subunit of the mammalian NC2 transcription inhibitor (20, 21). This mutation was already present in strain NG14 and may have been an early event in desensitizing the Pol II complex from negative regulation.

In addition to the high number of affected transcription factors, we also detected SNVs in three genes of RUT C30 that are involved in nuclear transport processes: Importin-beta 3 (tre78158), which is necessary to bind to the nuclear pore and thus target the importin-alpha-protein complex (22); nuclear transport factor 2 (tre22294), an essential component for the small GTPase Ran, which plays a central role in nucleocytoplasmic transport and which is hypothesized to exit the nucleus complexed with importin-beta (23); and an orthologue of ataxin-7 (tre112346), a protein essential for "gating" proteins to the nucleopore complex (24). These mutations may indicate that import of transcription factors and signaling proteins such as the MAP kinases (25), potentially required for transcription of cellulase genes, may be limiting at this stage.

Two of the mutations present in strain NG14 might affect the stability of mRNA: one of these genes encodes CCR4-associated factor 1 (CAF1), one of the two components in the well-characterized protein complex CCR4-NOT (tre110423), which mediates shortening of the poly (A) tail at the 3′ end of the mRNA (26). Data from plants suggest roles of this complex in regulated mRNA deadenylation and defense responses to pathogen infections (27). The second mutation potentially affecting RNA stability is a gene that encodes a component of the exosome (tre66895), a protein complex involved in maintaining correct RNA levels in eukaryotic cells (28). These findings are consistent with the observations that the mRNAs of secretory proteins are abundant for a longer time period in RUT C30 than in QM6a and suggest mRNA turnover as a target for improving extracellular protein production in this fungus.

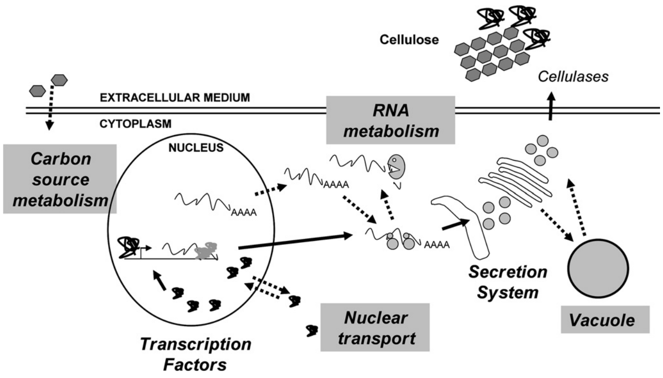

Fig. 5. Schematic "in context" representation of putative cellular processes affected in the strains used in this study. Black color indicates processes that where expected to be involved based on previous studies. Dotted arrows and gray squared text indicates potential areas of research for enhancement of cellulase secretion in *Trichoderma reesei* and other organisms.

Three of the six observed SNVs which concerned proteins involved in vesicle transport and secretion represented components necessary for vacuolar sorting (i.e., vacuolar sorting associated proteins VPS1 (tre43599) and VSP13 (tre65104), and the 16-kDa proteolipid subunit of the vacuolar ATPase (tre79014). These might suggest that the vacuole is a potential bottleneck for protein secretion in *T. reesei*. Interestingly, ultrastructural studies of xylanase II secretion by *T. reesei* demonstrated a considerable portion of the enzyme is localized in vacuoles (29, 30). An implication of the vacuoles as components of an alternative secretory pathway has been postulated for the slime mould *Dictyostelium discoideum* and the basidiomycete *Phanerochaete chrysosporium* (31, 32). In the former case, mannose-6-sulfate residues have been shown to target cathepsin D into the vacuolar route (33). Consistent with a role of the vacuoles in protein secretion, Harrison et al. (34) reported that the mannose chains of exocellobiohydrolase I (CEL7A) in a hyperproducing strain of *T. reesei* are sulfated. Although these bits of evidence are circumstantial, they make the role of the vacuoles in the generation of hyperproducing strains an interesting target for future research.

The high number of genes involved in metabolic pathways prompted us to examine carbon assimilation patterns of the strain on Biolog Phenotype microarrays. While these patterns coincide with some of the mutations observed (maltose metabolism in particular), it is interesting that these strains are affected in their rate of lactose metabolism. It is possible that this might lead to an enhanced pool of inducer in these strains. Since lactose is an inducer of cellulase production, it suggests that these strains, while already good cellulase producers on cellulosic material, are not fully optimized for producing cellulases on lactose.

Overall, our results build a strong genomic foundation on which we can build and test a large number of intriguing hypotheses about the mechanisms underlying *T. reesei* protein secretion, carbon catabolite repression and cellulose induction of cellulases. Finally, the results highlight so far neglected areas of research, such as nucleocytoplasmic transport, vacuolar protein trafficking and mRNA turnover for directed strain improvement (Fig. 5). These findings map a path forward toward identification of target genes in that when manipulated could greatly accelerate the development of improved industrial strains that are both safe and reliable for production of biofuels and biochemicals that are currently derived from non-renewable resources.

Materials and Methods

Massively Parallel Sequencing. Chromosomal DNA from *T. reesei* NG14 and RUT C30 were prepared as described previously (15). For samples that were single-end sequenced fragment libraries were prepared according to the

APPLIED BIOLOGICAL SCIENCES

genomic DNA sample prep protocol from Ilumina using 2.97 μg for NG14 and 1.5 μg for RUT C30. These libraries were then loaded onto the cluster generation station for single molecule bridge amplification using the Standard Cluster Generation kit from Ilumina. The slide with amplified clusters was then subjected to sequencing on the Ilumina Genome Analyser I (GAI) for single reads using the 36 cycle Sequencing Kit version 1 from Ilumina. We used four lanes for NG14 and four lanes for RUT C30 on the same flowcell. For the sample that was pair-end sequenced, fragment libraries were prepared according to the Ilumina Mate-Pair Library Prep kit protocol using 10.6 μg gDNA. This library was then loaded onto the cluster generation station for single molecule bridge amplification using the Paired-End Cluster Generation kit from Ilumina. The flowcell with amplified clusters was then subjected to sequencing on the Ilumina Genome Analyzer II (GAii) for paired-reads using 36 cycles in each direction with 36-cycle Sequencing version 2 kits from Ilumina. Four lanes were used for RUT C30 on the same flowcell, generating 3.9 to 6.7 million reads per lane. As a final control 120 randomly chosen SNVs were checked through Sanger sequencing in the three strains QM6a, NG14, and RUT C30.

Sequence Alignment and Analyses. We downloaded the QM6a strain reference genome v.2.0 from the Department of Energy Joint Genome Institute website (JGI): http://genome.jgi-psf.org/Trire2/. Solexa/Ilumina short reads from NG14 and RUT C30 strains were mapped onto the *T. reesei* genome using the Maq 0.6.6 software solution (35). Mapping was done with two maximum mismatches. SNVs where sorted from the consensus sequences obtained for each strain using cns2fq and cns2snp Maq functions. A first filtering step was done to discard SNVs with a read depth lower than 3, a mapping quality of reads greater than 60 and a Phred-like consensus quality lower than 30. Only exact SNVs (A, C, G, or T) were kept for further analyses.

We calculated GC percent, complexicity, and uniqueness scores by moving along a 60-bp window in the *T. reesei* genome. Complexity was calculated using the masked genome of *T. reesei* from the JGI website by counting the number of masked bases in each window. The uniqueness score was calculated using the genome tools from ArrayDesign (36) with the parameter MAX_PREFIX_LENGTH set to 30.

We evaluated the location of SNVs and deletions according to gene annotations using the ''filtered models'' from the JGI website. From this annotation we calculated the position of intron, promoter (using a 1-kb upstream region) and terminator (using a 200-base downstream region).

To assess the saturation of SNV identification, we simulated SNVs by altering the reference at known locations, aligning the data to this altered reference, and calculating the percentage of simulated SNVs that we were able locate. For the RUT C30 isolate that had paired-end data, we were also able to simulate short insertions and deletions (indels).

Biolog Phenotype Microarray Analysis. Global carbon assimilation patterns were investigated using Biolog FF MicroPlate™ (Biolog Inc.) using the protocol of Druzhinina et al. (37). Briefly, *T. reesei* strains were pregrown on 20 g/L malt extract agar, and 90 μL conidial suspension from them (75 ± 2% transmission at 590 nm) was dispensed into each of the wells of a Biolog FF MicroPlate™ (Biolog Inc.). The microplates were incubated in the dark at 30 °C, and percent absorbance determined after 12, 18, 24, 36, 42, 48, 66, and 72 h at 750 nm. Analyses were repeated at least three times for each strain. Basic statistical methods such as multiple regression analysis and analysis of variance (ANOVA) as well as multivariate exploratory techniques (cluster and factor analyses) were applied using Statistica 6.1 (StatSoft, Inc.) data analysis software system.

ACKNOWLEDGMENTS. The *H. jecorina*/*T. reesei* mutant genome sequencing by the Joint Genome Institute was funded by the United States Department of Energy (DOE) Office of Biological and Environmental Research, and accounted for one strain sequencing. Work in the Austrian lab was supported by a grant from the Austrian Science Foundation (FWF P-19160) to C.P.K. This work was supported by the Réseau National Génopôle (RNG) and IBISA for SLC. We thank Julie Poulain, Laurie Bertand, and Corinne Cruaud from the Genoscope/CNS for their help conducting the Solexa/Illumina sequencing for two strains. We thank Corinne Blugeon for SNV sequence validation, and Hugo Valls and Maya Spichal for their preliminary work on deletion detection. Analysis performed at Pacific Northwest National Laboratory was funded by the DOE Office of the Biomass Program.

1. Smith DR, et al. (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res* 18:1638–1642.
2. Bouws H, Wattenberg A, Zorn H (2008) Fungal secretomes–nature's toolbox for white biotechnology. *Appl Microbiol Biotechnol* 80:381–388.
3. Kumar R, Singh S, Singh OV (2008) Bioconversion of lignocellulosic biomass: Biochemical and molecular perspectives. *J Ind Microbiol Biotechnol* 35:377–391.
4. Stricker AR, Mach RL, de Graaff LH (2008) Regulation of transcription of cellulases- and hemicellulases-encoding genes in *Aspergillus niger* and *Hypocrea jecorina* (*Trichoderma reesei*). *Appl Microbiol Biotechnol* 78:211–220.
5. Reese ET (1976) History of the cellulase program at the U.S. army Natick Development Center. *Biotechnol Bioeng Symp* 6:9–20.
6. Eveleigh DE Montenecourt BS (1979) Increasing yields of extracellular enzymes. *Adv Appl Microbiol* 25:57–74.
7. Wick AN, Drury DR, Nakada HI, Wolfe JB (1957) Localization of the primary metabolic block produced by 2-deoxyglucose. *J Biol Chem* 224:963–969.
8. Kang HT, Hwang ES (2006) 2-Deoxyglucose: An anticancer and antiviral therapeutic, but not anymore a low glucose mimetic. *Life Sci* 78:1392–1399.
9. Herpoel-Gimbert I, et al. (2008) Comparative secretome analyses of two *Trichoderma reesei* RUT-C30 and CL847 hypersecretory strains. *Biotechnol Biofuels* 1:18.
10. Pakula TM, Salonen K, Uusitalo J, Penttila M (2005) The effect of specific growth rate on protein synthesis and secretion in the filamentous fungus *Trichoderma reesei*. *Microbiology* 151:135–143.
11. Carter GL, Allison D, Rey MW, Dunn-Coleman NS (1992) Chromosomal and genetic analysis of the electrophoretic karyotype of *Trichoderma reesei*: Mapping of the cellulase and xylanase genes. *Mol Microbiol* 6:2167–2174.
12. Mantyla AL, et al. (1992) Electrophoretic karyotyping of wild-type and mutant *Trichoderma longibrachiatum* (*reesei*) strains. *Curr Genet* 21:471–477.
13. Ilmen M, Thrane C, Penttila M (1996) The glucose repressor gene cre1 of *Trichoderma*: Isolation and expression of a full-length and a truncated mutant form. *Mol Gen Genet* 251:451–460.
14. Geysens S, et al. (2005) Cloning and characterization of the glucosidase II alpha subunit gene of *Trichoderma reesei*: A frameshift mutation results in the aberrant glycosylation profile of the hypercellulolytic strain Rut-C30. *Appl Environ Microbiol* 71:2910–2924.
15. Seidl V, et al. (2008) The *Hypocrea jecorina* (*Trichoderma reesei*) hypercellulolytic mutant RUT C30 lacks a 85 kb (29 gene-encoding) region of the wild-type genome. *BMC Genomics* 9:327.
16. Martinez D, et al. (2008) Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). *Nat Biotechnol* 26:553–560.
17. Ohnishi J, Mizoguchi H, Takeno S, Ikeda M (2008) Characterization of mutations induced by N-methyl-N′-nitro-N-nitrosoguanidine in an industrial *Corynebacterium glutamicum* strain. *Mutat Res* 649:239–244.
18. Ralser M, et al. (2008) A catabolic block does not sufficiently explain how 2-deoxy-D-glucose inhibits cell growth. *Proc Natl Acad Sci USA* 105:17807–17811.
19. Todd RB, Andrianopoulos A (1997) Evolution of a fungal regulatory gene family: The Zn(II)2Cys6 binuclear cluster DNA binding motif. *Fungal Genet Biol* 21:388–405.
20. Albert TK, et al. (2007) Global distribution of negative cofactor 2 subunit-alpha on human promoters. *Proc Natl Acad Sci USA* 104:10000–10005.
21. Kim TK, Zhao Y, Ge H, Bernstein R, Roeder RG (1995) TATA-binding protein residues implicated in a functional interplay between negative cofactor NC2 (Dr1) and general factors TFIIA and TFIIB. *J Biol Chem* 270:10976–10981.
22. Cook A, Bono F, Jinek M, Conti E (2007) Structural biology of nucleocytoplasmic transport. *Annu Rev Biochem* 76:647–671.
23. George R, et al. (2009) A complex of Shc and Ran-GTPase localises to the cell nucleus. *Cell Mol Life Sci* 66:711–720.
24. Kohler A, Schneider M, Cabal GG, Nehrbass U, Hurt E (2008) Yeast ataxin-7 links histone deubiquitination with gene gating and mRNA export. *Nat Cell Biol* 10:707–715.
25. James BP, Bunch TA, Krishnamoorthy S, Perkins LA, Brower DL (2007) Nuclear localization of the ERK MAP kinase mediated by Drosophila alphaPS2-betaPS integrin and importin-7. *Mol Biol Cell* 18:4190–4199.
26. Garapaty S, Mahajan MA, Samuels HH (2008) Components of the CCR4-NOT complex function as nuclear hormone receptor coactivators via association with the NRC-interacting Factor NIF-1. *J Biol Chem* 283:6806–6816.
27. Liang W, et al. (2008) The Arabidopsis homologs of CCR4-associated factor 1 show mRNA deadenylation activity and play a role in plant defense responses. *Cell Res* 19:307–316.
28. Raijmakers R, Schilders G, Pruijn GJ (2004) The exosome, a molecular machine for controlled RNA degradation in both nucleus and cytoplasm. *Eur J Cell Biol* 83:175–183.
29. Kurzatkowski W, et al. (1993) Ultrastructural localization of cellular compartments involved in secretion of the low molecular weight, alkaline xylanase by *Trichoderma reesei*. *Arch Microbiol* 159:417–422.
30. Torronen A, Rouvinen J (1995) Structural comparison of two major endo-1,4-xylanases from *Trichoderma reesei*. *Biochemistry* 34:847–856.
31. Kuan IC, Tien M (1989) Phosphorylation of lignin peroxidases from *Phanerochaete chrysosporium*. Identification of mannose 6-phosphate. *J Biol Chem* 264:20350–20355.
32. Souza GM, et al. (1997) *Dictyostelium* lysosomal proteins with different sugar modifications sort to functionally distinct compartments. *J Cell Sci* 110:2239–2248.
33. Journet A, et al. (1999) Characterization of *Dictyostelium discoideum* cathepsin D. *J Cell Sci* 112:3833–3843.
34. Harrison, et al. (1998) Modified glycosylation of cellobiohydrolase I from a high cellulase-producing mutant strain of Trichoderma reesei. *Eur J Biochem* 256:119–127.
35. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851–1858.
36. Graf S, et al. (2007) Optimized design and assessment of whole genome tiling arrays. *Bioinformatics* 23:i195–204.
37. Druzhinina IS, Schmoll M, Seiboth B, Kubicek CP (2006) Global carbon utilization profiles of wild-type, mutant, and transformant strains of Hypocrea jecorina. *Appl Environ Microbiol* 72:2126–2133.