

Vergleichende Genomanalyse

Ist das Schimpansen-Genom sequenziert?

STEFAN TAUDIEN¹, INGO EBERSBERGER², NIELS JAHN¹, MATTHIAS PLATZER¹

¹LEIBNIZ-INSTITUT FÜR ALTERSFORSCHUNG, FRITZ-LIPMANN-INSTITUT, JENA,

²ZENTRUM FÜR INTEGRATIVE BIOINFORMATIK, MAX F. PERUTZ LABORATORIEN, WIEN

Die Sequenzierung des menschlichen Genoms ist ein Meilenstein der bio-wissenschaftlichen Forschung. Die Abfolge von ca. 3 Mrd. Basenpaaren beinhaltet die Information, das System „Mensch“ zu begreifen. Welche Gene und sonstigen funktionellen Bereiche kodiert sind und welche Aufgabe sie haben, sind vordringliche Fragestellungen der Genomanalyse. Der Vergleich von Genomsequenzen zwischen Arten liefert wichtige Hinweise.

Vergleichende Genomanalyse

■ Regionen, deren DNA-Sequenz über weite Distanzen konserviert geblieben ist, sind vielversprechende Kandidaten bei der Suche. Deswegen wurde in den vergangenen Jahren weltweit die Sequenzierung vieler eukaryotischer Genome initiiert. Trotz der technologischen Fortschritte ist die Bestimmung einer Genomsequenz mit der gleichen Vollständigkeit und hohen Qualität wie beim Menschen immer noch enorm zeit- und kostenintensiv. Die Erstellung einer Rohversion (*working draft*, WD) durch genomweite Sequenzierung von Zufallsfragmenten (*whole genome shotgun*, WGS), die große Teile des Genoms mit guter bis sehr guter Qualität abdeckt, macht dabei nur einen geringen Anteil aus. Für viele Fragestellungen reicht ein WD aus. Entsprechend sehen viele Genomprojekte eine abschließende Vervollständigung der Sequenz (*finishing*) nicht vor. Dieser Verzicht ist im Falle des Schimpansengenoms problematisch.

Das Schimpansengenom – ein „besonderer Fall“

Der Schimpanse (*Pan troglodytes*) ist unser nächster lebender Verwandter. Sein Genom unterscheidet sich von dem des Menschen in nur ~1,2 %^[1–3]. Diese Zahl sinkt noch einmal um etwa die Hälfte, wenn ausschließlich proteinkodierende Sequenzen verglichen werden^[4, 5]. Der genetische Unterschied ist damit

nur etwa zehnfach höher als zwischen zwei zufällig ausgesuchten Menschen^[6]. Unter den vergleichsweise wenigen genetischen Unterschieden zwischen Mensch und Schimpanse sollten sich diejenigen identifizieren lassen, die für „typisch menschliche“ Merkmale verantwortlich sind. Weiterhin hofft man Gene zu identifizieren, die in einer der beiden Arten mehr proteinverändernde Nukleotidaustausche angesammelt haben als erwartet. Solche Gene könnten in der betroffenen Art an Bedeutung verloren haben und einem geringeren Selektionsdruck ausgesetzt sein^[7]. Alternativ kann es sich auch um Gene handeln, die unter positiver Selektion stehen^[8–10] (**Abb. 1**). Vom Vergleich mit entfernteren Arten zum Vergleich Mensch-Schimpanse hat folglich ein Paradigmenwechsel stattgefunden. Nun machen die kleinen Unterschiede die gewünschte Information aus, nicht länger die Sequenzähnlichkeit. Darum kommt der Genauigkeit der Schimpansengenomsequenz eine besondere Bedeutung zu. Rechtfertigt das den großen Aufwand der Überführung des WD in eine dem Humangenom qualitativ gleichwertige „finale“ Sequenz?

Kleine Ursache ...

Der Vorteil einer „finalen“ Schimpansensequenz lässt sich am besten abschätzen, wenn man die gleiche Analyse mit beiden Varianten durchführt und die Ergebnisse vergleicht. Dafür hatten wir 2005 einen Datensatz erstellt

(ptrHQ), der in seiner Sequenzqualität der humanen Genomsequenz entspricht (172 kb; Fehlerrate: < 1/10.000 bp; *Coverage*: 8–10)^[11]. Die verwendeten Sequenzen umfassen 653 Exons aus 54 Genen von den Chromosomen 21 und X. 137 kb (79 %) waren durch entsprechende Sequenzen im Schimpansen-WD (panTro1) vertreten und bildeten die Grundlage der Analyse. Die Abdeckung der X-chromosomalen Bereiche durch die WD-Sequenz war mit 78 % geringer als die der Regionen von Chromosom 21 (84 %). Dieser Unterschied erklärt sich aus der WGS-Sequenzierungsstrategie und hat seine Ursache darin, dass die WD-Sequenzen dem Genom eines männlichen Schimpansen entstammen, das im Gegensatz zu zwei Kopien jedes Autosoms nur eine des X-Chromosoms enthält.

Der direkte Vergleich der beiden Schimpansen-Datensätze (ptrHQ und ptrWD) ergab Unterschiede in 425 Positionen (0,3 %). Ein Teil kann auf allelische Variationen zurückgeführt werden, da die Sequenzen von unterschiedlichen Individuen stammen. Der Wert von 0,3 % liegt allerdings deutlich über den erwarteten 0,08–0,13 %, die in früheren Studien als mittlerer Sequenzunterschied zwischen Schimpansen-Individuen ermittelt worden sind^[12, 13]. Somit liegt nahe, für einen Großteil Sequenzierfehler im ptrWD-Datensatz als Ursache anzunehmen. Tatsächlich weisen 328 der 425 differierenden Positionen (77 %) im WD niedrige Qualität ($Q^* < 40$, d. h. Fehlerrate: > 1/10.000 bp) auf.

Der Vergleich der beiden Schimpansen- mit den entsprechenden menschlichen Exonsequenzen liefert weitere Hinweise auf Sequenzierfehler im ptrWD. Bei einem Sequenzunterschied von ca. 1 % zwischen Mensch und

* Q (Qualitäts- bzw. Konfidenzwert) – Maß für die Sequenzqualität, Grad der Übereinstimmung zwischen experimentell bestimmter und tatsächlicher Basenabfolge. Der Q -Wert berechnet sich aus der Fehlerwahrscheinlichkeit P_e einer entsprechenden Position in der experimentell bestimmten Sequenz gemäß $Q = -10 \cdot \log_{10}(P_e)$. $Q = 40$ entspricht 1 Fehler/10.000 bp bzw. einer Genauigkeit von 99,99 %, $Q = 20$, 1 Fehler/100 bp bzw. 99 %.

Schimpanse bzw. ca. 0,1 % zwischen zwei Schimpansen liegt die Wahrscheinlichkeit, dass eine Position, in der sich beide Arten unterscheiden, gleichzeitig polymorph bei den Schimpansen ist, in der Größenordnung von $0,01 \times 0,001 = 10^{-5}$ (1/100.000 bp). In einem Vergleich von 137 kb sollten daher nahezu alle Basenaustausche zwischen Mensch und Schimpanse beobachtet werden, unabhängig von der herangezogenen Schimpansensequenz. Der ptrHQ-Datensatz weist 436 Unterschiede zur menschlichen Sequenz auf (Tab. 1), die alle auch im ptrWD-Vergleich erfasst werden. Allelische Variationen der Schimpansensequenzen sind in dieser Analyse daher tatsächlich vernachlässigbar. Im Vergleich ptrWD-human werden hingegen 103 zusätzliche Unterschiede gefunden, die durch die ptrHQ-Daten nicht unterstützt werden.

... große Wirkung

Diese 103 Differenzen stellen vermutlich Sequenzierfehler im WD dar. Dies entspräche einer Fehlerrate von $< 1/1.000$ bp – eine Größenordnung, die den Standardanforderungen vergleichender Genomanalysen genügt. Im vorliegenden Vergleich proteinkodierender Sequenzen zwischen Mensch und Schimpanse hat diese kleine Fehlerrate jedoch weitreichende Konsequenzen: Engt man die Betrachtung auf Nukleotidaustausche ein, die einen Aminosäureaustausch zur Folge hätten, erhöht sich der Anteil an Unterschieden aufgrund von Sequenzierfehlern („falsch Positiv“) von 19 % (103/539) auf 25 %. Dies bewirkt, dass einzelne Gene Hinweise auf positive Selektion liefern, die sich mit den ptrHQ-Daten nicht bestätigen lassen.

Trotz dieser offensichtlichen Problematik gibt es eine Möglichkeit, WD-Sequenzen für solche Analysen heranzuziehen, falls die Sequenzqualität berück-

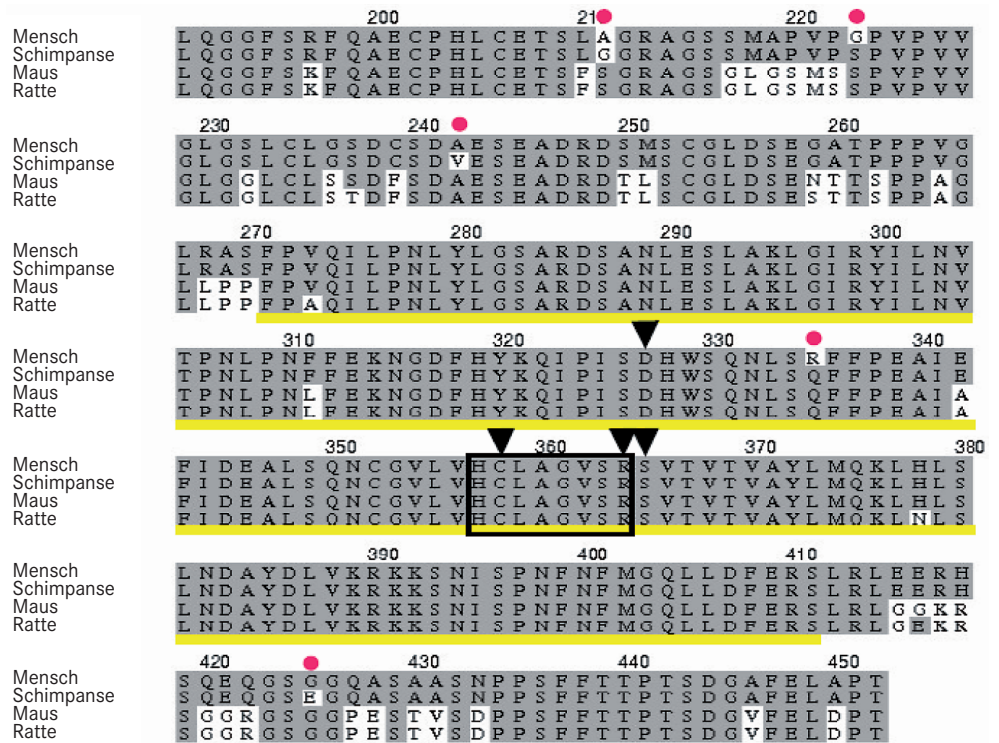
sichtigt wird. 75 % der 103 vermutlich falsch positiven Differenzen fallen auf Positionen im WD mit $Q < 40$ (65 %, $Q < 20$). Berücksichtigt man nur Positionen mit $Q \geq 40$, erhöht man die Spezifität der Analyse von 81 % auf 93 % (Tab. 1), d. h. 93 % der gefundenen Mensch-Schimpanse-Unterschiede werden durch den ptrHQ-Datensatz unterstützt. Dieser Gewinn geht allerdings auf Kosten der Sensitivität. Nur 86 % der Unterschiede werden noch gefunden, da die verbleibenden Positionen von der Analyse ausgeschlossen werden. Eine Limitierung auf $Q \geq 20$ hingegen liefert bei vergleichbarer Spezifität (92 %) immer noch eine Sensitivität von 96 %.

Initiativen zur Verbesserung des Schimpansen-WD

Unvollständige Sequenzabdeckung, Sequenzier-

fehler sowie nicht aufzulösende Gentruckturen in duplizierten Bereichen (Abb. 2) sind Probleme aller WDs. Unsere Studie zeigt, dass WD-Sequenzen in der komparativen Analyse eng verwandter Spezies wie Mensch und Schimpanse dennoch zufriedenstellende Resultate erbringen können, wenn Qualitätskriterien angemessen berücksichtigt werden. Ein zuverlässiger und voll-

ständiger Katalog der genetischen Differenzen ist jedoch nur mit einer „finalen“ Sequenzassemblierung zu erstellen. Seit der Originalpublikation des Schimpansengenoms (panTro1^[3]) ist dank des Einsatzes des Genom-Sequenzierzentrums der Washington University (St. Louis, USA) die Abdeckung des WD von 3,6 auf 6 angehoben worden (panTro2). Der resultierende Anstieg



▲ **Abb. 1:** Proteinsequenzen der dualspezifischen Phosphatase DUSP9. Vor allem die katalytische Domäne ist stark konserviert (gelbe Linie, funktionelle Aminosäuren durch Pfeile, aktives Zentrum durch Rahmen markiert). Von insgesamt sieben Nukleotidunterschieden Mensch-Schimpanse (ptrHQ) führen fünf zu Aminosäureaustauschen (rote Punkte) und deuten auf positive Selektion hin. Vier der Austausche liegen außerhalb, ein weiterer innerhalb der katalytischen Domäne.

NEW

Customised complete
monoclonal antibodies

- 200 µg antigen
- 3–4 months
- customer's property

BioGENES

Berlin • Germany • www.biogenes.de • +49 (0)30-65 76 23 96

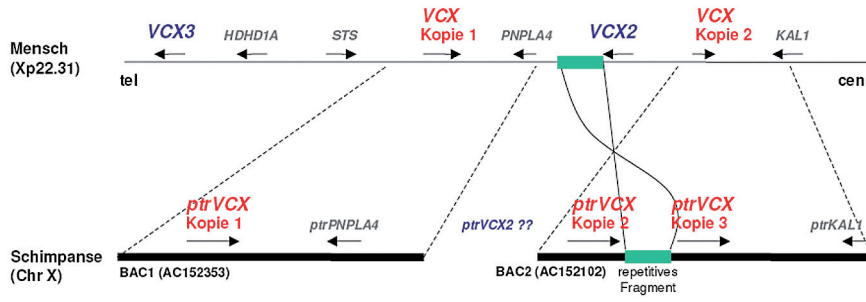


Abb. 2: Der humane VCX-Locus enthält drei sehr ähnliche, paraloge Gene. Derartig komplexe Strukturen kann eine WGS-Strategie im Gegensatz zu klonbasierter Sequenzierung *a priori* nicht auflösen. Die beiden dargestellten Schimpansen-BACs (ptrHQ) beinhalten drei Kopien des zu VCX orthologen Gens (*ptrVCX*), wobei die flankierenden Gene (*ptrPNPLA4* bzw. *ptrKAL1*) die eindeutige Zuordnung zum humanen Locus ermöglichen. Die Nukleotidentität zwischen den *ptrVCX*-Kopien beträgt 99 %, zwischen den humanen Paralogen VCX, VCX2 und VCX3 hingegen 94–98 %, sodass die beiden auf BAC2 gelegenen Gene tatsächlich als Kopien von *ptrVCX* angesehen werden müssen. Ein zu VCX2 orthologes Gen ist daher auf einem Klon zu erwarten, der BAC1 und BAC2 verbindet.

Nukleotiddifferenzen ^a Mensch-Schimpanse	gemeinsam ^b	unikal ^c	insgesamt	Spezifität ^d	Sensitivität ^e
alle Positionen	436	103	539	81	% entspricht 100 %
Q ≥ 20	418	36	454	92 %	96 %
Q ≥ 40	374	26	400	93 %	86 %

^a bezogen auf 136.547 im Schimpansen-WD enthaltene Positionen des Datensatzes

^b Differenzen Mensch-Schimpanse, die sowohl durch finale Sequenz (ptrHQ) als auch durch Sequenzen der Arbeitsversion (ptrWD) repräsentiert werden

^c Differenzen Mensch-Schimpanse, die ausschließlich durch Sequenzen der Arbeitsversion (ptrWD) repräsentiert werden

^d gemeinsam/gesamt

^e gemeinsam/436

Tab. 1: Identifikation echter Nukleotidunterschiede zwischen Mensch und Schimpanse unter Berücksichtigung von Schimpansen-WD-Positionen mit niedriger bzw. mittlerer Sequenzqualität.

der Sequenzqualität zusammen mit der Abdeckung von mittlerweile ca. 97 % des euchromatischen Teils des Schimpansen-genoms stellt eine profunde Datenbasis dar. Dabei ist das durch ein internationales Konsortium unter Beteiligung von drei deutschen Sequenzierungszentren analysierte Schimpansen-Chromosom 21 das einzige Autosom in finaler Qualität^[14]. Eine entsprechende Sequenz der pseudoautosomalen Anteile des Y-Chromosoms wurde in zwei Projekten parallel bestimmt^[15, 16]. In beiden Fällen wurde das Chromosom eines Vertreters der westafrikanischen Unterart analysiert, sodass die unabhängig ermittelten Ergebnisse auf systematische Fehler hin untersucht werden können. Andererseits wurde durch mangelnde Abstimmung die Chance verpasst, am Beispiel des Y die Chromosomen von zwei Unterarten zu vergleichen. Ein vom deutschen Nationalen Genomforschungsnetz (NGFN2) gefördertes Konsortium (MPIMG Berlin, FLI Jena, HZI Braunschweig) arbeitet seit 2005 an der Überführung von Teilen des X-Chromosom-WDs in die finale Sequenz (www.

science.ngfn.de/10_440.htm). Dabei werden BAC-Klone anhand ihrer Endsequenzen *in silico* kartiert und zur Erhöhung der lokalen Abdeckung niedrig-redundant sequenziert. Mit der für Mitte 2008 geplanten Fertigstellung wird dann 40 Mb hochwertige Sequenz u. a. sehr genreicher Regionen sowie von etlichen Genen mit Assoziation zu X-gekoppelter mentaler Retardation zur Verfügung stehen. Es bleibt zu hoffen, dass in Zukunft das gesamte Schimpansen-genom schrittweise, möglicherweise durch den Einsatz neuer Sequenzierungstechnologien, in finale Qualität überführt werden kann. Dies würde uns dem Verständnis dessen, was uns zu Menschen macht, einen entscheidenden Schritt näher bringen.

Literatur

- [1] Ebersberger, I., Metzler, D., Schwarz, C., Paabo, S. (2002): Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* 70: 1490–1497.
- [2] Fujiyama, A., Watanabe, H., Toyoda, A., Taylor, T. D., Itoh, T., Tsai, S. F., Park, H. S., Yaspo, M. L., Lehrach, H., Chen, Z., Fu, G., Saitou, N., Osoegawa, K., de Jong, P. J., Suto, Y., Hattori, M., Sakaki, Y. (2002): Construction and analysis of a

human-chimpanzee comparative clone map. *Science* 295: 131–134.

[3] The Chimpanzee Sequencing Consortium (2005): Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437: 69–87.

[4] Hellmann, I., Zollner, S., Enard, W., Ebersberger, I., Nickel, B., Paabo, S. (2003): Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* 13: 831–837.

[5] Shi, J., Xi, H., Wang, Y., Zhang, C., Jiang, Z., Zhang, K., Shen, Y., Jin, L., Yuan, W., Lin, J., Hua, Q., Wang, F., Xu, S., Ren, S., Zhao, G., Chen, Z., Huang, W. (2003): Divergence of the genes on human chromosome 21 between human and other hominoids and variation of substitution rates among transcription units. *Proc. Natl. Acad. Sci. USA* 100: 8331–8336.

[6] Crawford, D. C., Akey, D. T., Nickerson, D. A. (2005): The patterns of natural variation in human genes. *Annu. Rev. Genomics Hum. Genet.* 6: 287–312.

[7] Gilad, Y., Man, O., Glusman, G. (2005): A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res.* 15: 224–230.

[8] Arbiza, L., Dopazo, J., Dopazo, H. (2006): Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput. Biol.* 2: e38.

[9] Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., Civello, D., Adams, M. D., Cargill, M., Clark, A. G. (2005): Natural selection on protein-coding genes in the human genome. *Nature* 437: 1153–1157.

[10] Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., Ferriera, S., Wang, G., Zheng, X., White, T. J., Sninsky, J. J., Adams, M. D., Cargill, M. (2003): Inferring non-neutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302: 1960–1963.

[11] Taudien, S., Ebersberger, I., Glockner, G., Platzer, M. (2006): Should the draft chimpanzee sequence be finished? *Trends Genet.* 22: 122–125.

[12] Kaessmann, H., Wiebe, V., Paabo, S. (1999): Extensive nuclear DNA sequence diversity among chimpanzees. *Science* 286: 1159–1162.

[13] Yu, N., Jensen-Seaman, M. I., Chernick, L., Kidd, J. R., Deinard, A. S., Ryder, O., Kidd, K. K., Li, W. H. (2003): Low nucleotide diversity in chimpanzees and bonobos. *Genetics* 164: 1511–1518.

[14] Watanabe, H., Fujiyama, A., Hattori, M., Taylor, T. D., Toyoda, A., Kuroki, Y., Noguchi, H., BenKahla, A., Lehrach, H., Sudbrak, R., Kube, M., Taenzler, S., Galgoczy, P., Platzer, M., Scharfe, M., Nordsiek, G., Blocker, H., Hellmann, I., Khaitovich, P., Paabo, S., Reinhardt, R., Zheng, H. J., Zhang, X. L., Zhu, G. F., Wang, B. F., Fu, G., Ren, S. X., Zhao, G. P., Chen, Z., Lee, Y. S., Cheong, J. E., Choi, S. H., Wu, K. M., Liu, T. T., Hsiao, K. J., Tsai, S. F., Kim, C. G., Oota, S., Kitano, T., Kohara, Y., Saitou, N., Park, H. S., Wang, S. Y., Yaspo, M. L., Sakaki, Y. (2004): DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* 429: 382–388.

[15] Hughes, J. F., Skaletsky, H., Pyntikova, T., Minx, P. J., Graves, T., Rozen, S., Wilson, R. K., Page, D. C. (2005): Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature* 437: 100–103.

[16] Kuroki, Y., Toyoda, A., Noguchi, H., Taylor, T. D., Itoh, T., Kim, D. S., Kim, D. W., Choi, S. H., Kim, I. C., Choi, H. H., Kim, Y. S., Satta, Y., Saitou, N., Yamada, T., Morishita, S., Hattori, M., Sakaki, Y., Park, H. S., Fujiyama, A. (2006): Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nat. Genet.* 38: 158–167.

Korrespondenzadresse:



Dr. Matthias Platzer
Genomanalyse
Leibniz-Institut für Altersforschung
Fritz-Lipmann-Institut e. V.
Beutenbergstr. 11
D-07745 Jena
Tel.: 03641-656241
Fax: 03641-656255
mplatzer@fli-leibniz.de
http://genome.fli-leibniz.de